

# GPT-5.5 annotation evaluation — ten attempts on each of four tasks

4 fixed tasks · 40 sealed attempt records · 40  
graded attempts · 25 certified passes

CURRENT  
CAMPAIGN REPORT  
2026-09-06

Evaluation scope:  
cli/codex-gpt-5.5

Prepared by:  
VVDex Forge — regenerated from verified sealed records only

fc-ec21239c00aa · 40 record(s) · every value recomputable

# Contents

|     |                                            |     |
|-----|--------------------------------------------|-----|
| 01. | Executive summary                          | 4   |
| 02. | Per-exam results                           | 5   |
| 03. | Environment families                       | 6   |
| 04. | Campaign aggregate across these four exams | 7   |
| 05. | Harness status                             | 8   |
| 06. | Cells excluded from capability             | 9   |
| 07. | What the outcomes mean                     | 10  |
| 08. | What this report withholds, and why        | 11  |
| 09. | Attempt record 1 of 10 · image-object-1    | 12  |
| 10. | Attempt record 2 of 10 · image-object-1    | 27  |
| 11. | Attempt record 3 of 10 · image-object-1    | 42  |
| 12. | Attempt record 4 of 10 · image-object-1    | 57  |
| 13. | Attempt record 5 of 10 · image-object-1    | 72  |
| 14. | Attempt record 6 of 10 · image-object-1    | 87  |
| 15. | Attempt record 7 of 10 · image-object-1    | 102 |
| 16. | Attempt record 8 of 10 · image-object-1    | 117 |
| 17. | Attempt record 9 of 10 · image-object-1    | 132 |
| 18. | Attempt record 10 of 10 · image-object-1   | 147 |
| 19. | Attempt record 1 of 10 · robot-video-1     | 162 |
| 20. | Attempt record 2 of 10 · robot-video-1     | 176 |
| 21. | Attempt record 3 of 10 · robot-video-1     | 190 |
| 22. | Attempt record 4 of 10 · robot-video-1     | 205 |
| 23. | Attempt record 5 of 10 · robot-video-1     | 220 |
| 24. | Attempt record 6 of 10 · robot-video-1     | 235 |
| 25. | Attempt record 7 of 10 · robot-video-1     | 249 |
| 26. | Attempt record 8 of 10 · robot-video-1     | 264 |
| 27. | Attempt record 9 of 10 · robot-video-1     | 278 |
| 28. | Attempt record 10 of 10 · robot-video-1    | 292 |
| 29. | Attempt record 1 of 10 · structured-data-1 | 307 |
| 30. | Attempt record 2 of 10 · structured-data-1 | 322 |
| 31. | Attempt record 3 of 10 · structured-data-1 | 337 |

32. Attempt record 4 of 10 · structured-data-1 ..... 352

33. Attempt record 5 of 10 · structured-data-1 ..... 367

34. Attempt record 6 of 10 · structured-data-1 ..... 382

35. Attempt record 7 of 10 · structured-data-1 ..... 397

36. Attempt record 8 of 10 · structured-data-1 ..... 412

37. Attempt record 9 of 10 · structured-data-1 ..... 427

38. Attempt record 10 of 10 · structured-data-1 ..... 442

39. Attempt record 1 of 10 · text-labeling-1 ..... 457

40. Attempt record 2 of 10 · text-labeling-1 ..... 472

41. Attempt record 3 of 10 · text-labeling-1 ..... 487

42. Attempt record 4 of 10 · text-labeling-1 ..... 502

43. Attempt record 5 of 10 · text-labeling-1 ..... 517

44. Attempt record 6 of 10 · text-labeling-1 ..... 532

45. Attempt record 7 of 10 · text-labeling-1 ..... 547

46. Attempt record 8 of 10 · text-labeling-1 ..... 562

47. Attempt record 9 of 10 · text-labeling-1 ..... 577

48. Attempt record 10 of 10 · text-labeling-1 ..... 592

49. Evidence ..... 607

50. Attestation ..... 609

01 Executive summary

**Release status: CURRENT.** This is the campaign the published featured index carries; its exams' public pages state these numbers.

|                                          |                                           |                                                  |                                            |
|------------------------------------------|-------------------------------------------|--------------------------------------------------|--------------------------------------------|
| <div>FIXED TASKS</div> <div>4</div>      | <div>SEALED RECORDS</div> <div>40</div>   | <div>MODEL LANES</div> <div>1</div>              | <div>ATTEMPTS</div> <div>40</div>          |
| <div>GRADED ATTEMPTS</div> <div>40</div> | <div>CERTIFIED PASSES</div> <div>25</div> | <div>LANE ERRORS (NOT GRADED)</div> <div>0</div> | <div>WITHHELD / INVALID</div> <div>0</div> |

Purpose

This report presents a controlled evaluation of 1 model lane(s) against 4 fixed certified annotation tasks, ten sealed attempt records on each task. For each task, every lane received that task's frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw; and every rollout carries three separate outcomes — grader, integrity, certified — of which only the certified outcome is counted. Each sealed attempt record is rendered below as a chapter derived from the same record as its standalone report: the chapter's bytes are authenticated by this campaign document, the standalone report by its own digest.

**Annotation campaign scope.** This campaign evaluates four fixed annotation tasks with ten sealed attempt records on each task (40 sealed attempts in total). The sealed tasks cover image, structured, text, video inputs. Video was evaluated through the sampled still frames recorded in every video attempt; these results do not establish native-video capability. Audio is N/A because no audio task appears in this campaign's sealed records. CLI temperature and stochastic-seed controls are not established by these records; attempt indices are bookkeeping. Token use and per-attempt cost are N/A because every recorded lane result marks token telemetry unavailable and no cost measurement is attached. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

25

certified pass(es) across 40 graded attempts (40 attempts, 0 lane error(s) not graded) in 4 cells — counted only where the grader accepted the submission AND containment was verified AND provenance is complete AND no integrity invariant or verdict-bearing harness suspicion applies.

fc-ec21239c00aa · generated from 40 verified sealed record(s)

Key findings across the campaign

- image-object-1 · cli/codex-gpt-5.5 · **ATTENTION** — 5 certified passes / 10 graded attempts; 5 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- robot-video-1 · cli/codex-gpt-5.5 · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · cli/codex-gpt-5.5 · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/codex-gpt-5.5 · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

| MODEL LANE | ANNOTATION · IMAGE-<br>OBJECT-1<br>VVDEX.ANNOTATION.IM<br>AGE-OBJECT-1 | ANNOTATION · ROBOT-<br>VIDEO-1<br>VVDEX.ANNOTATION.RO<br>BOT-VIDEO-1 | ANNOTATION · STRUCTURED-<br>DATA-1<br>VVDEX.ANNOTATION.STRU<br>CTURED-DATA-1 | ANNOTATION · TEXT-<br>LABELING-1<br>VVDEX.ANNOTATION.TEX<br>T-LABELING-1 |
|------------|------------------------------------------------------------------------|----------------------------------------------------------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------|
|------------|------------------------------------------------------------------------|----------------------------------------------------------------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------|

|                   |      |      |       |       |
|-------------------|------|------|-------|-------|
| cli/codex-gpt-5.5 | 5/10 | 0/10 | 10/10 | 10/10 |
|-------------------|------|------|-------|-------|

Each cell is *certified passes / graded attempts* for that lane on that exam. An attempt the lane ended before grading is reported beside the cell — 1/9 + 1 lane error, never 1/10 — and is outside the denominator; withheld and invalid attempts are reported the same way. A cell whose graded attempts all passed but which also holds such an attempt is shown as *passes with attempts outside the denominator*, not as a clean sweep — the clean-sweep state is reserved for a cell where every attempt was graded and every graded attempt certified. No cell here carries a percentage: these are separate exams, not one score.

03 Environment families

This campaign covers four exam(s) drawn from one environment family(ies). The sentence in each row is the family's own description of what the exam tests; nothing is claimed here that the exam does not.

| FAMILY                | EXAM                                            | WHAT IT TESTS                                                                                                   |
|-----------------------|-------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Multimodal annotation | <code>vvdex.annotation.image-object-1</code>    | A constructed image is reviewed for classification, object boxes, attributes and keypoints.                     |
| Multimodal annotation | <code>vvdex.annotation.robot-video-1</code>     | A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion. |
| Multimodal annotation | <code>vvdex.annotation.structured-data-1</code> | Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.     |
| Multimodal annotation | <code>vvdex.annotation.text-labeling-1</code>   | An invented message is reviewed for classification and entity spans with consistent offsets.                    |

04 Campaign aggregate across these four exams

| MODEL LANE        | CERTIFIED PASSES | GRADED ATTEMPTS | LANE ERRORS | WITHHELD | INVALID |
|-------------------|------------------|-----------------|-------------|----------|---------|
| cli/codex-gpt-5.5 | 25               | 40              | 0           | 0        | 0       |

Raw counts first. *Graded attempts* = certified passes + model fails; lane errors, withheld and invalid cells are outside capability arithmetic in both directions. These four fixed annotation tasks measure different work, so no combined rate or confidence interval is reported. The totals are accounting counts only; the per-task results above carry the quality claims.

## 05 Harness status

---

No record in this campaign fired a harness self-suspicion rule. Every countable cell is a grader outcome under verified containment.



## 06 Cells excluded from capability

---

An attempt is excluded from capability only when the lane, not the model, ended it (lane error), when the run cannot prove its own conditions (withheld), or when an integrity invariant fired (INVALID). A submission the hidden grader failed is a model result and is counted, never listed here.

No attempt in this campaign was excluded: every attempt produced a countable model result.

## 07 What the outcomes mean

- **CERTIFIED PASS:** the grader accepted the submission, containment was verified, provenance is complete and no integrity invariant fired.
- **model fail:** the submission failed the hidden tests under verified containment; a real, attributable model result.
- **lane error:** the provider or CLI lane failed (rate limit, quota, transport) before an attributable measurement existed; not a model result.
- **withheld:** the run cannot prove its own conditions (containment unproven, provenance incomplete, or a harness suspicion the record cannot clear); neither a pass nor a model failure.
- **INVALID:** an integrity invariant fired (isolation breach, evidence mismatch); the row is not a model result.
- **harness suspect:** the run's own self-check fired; each rule is classified as verdict-bearing or diagnostic, and verdict-bearing suspicion withholds the affected cells.

08 What this report withholds, and why

What a public document may print about a rollout is decided by what the exam GRADES, not case by case. Where the graded artefact is a change to a public upstream repository, a bounded redacted excerpt of the model's own diff is published. Where the graded output is an answer, a keyed value or the state of an application, nothing derived from the submission is published — no diff, no answer value, no model claim, and not the names of the hidden assertions a rollout failed, because a grader's name can carry the value it asserts. Withheld on every exam without exception: the contents of tool results, the hidden tests, the reference solution, host paths and keys. The counts, verdicts and intervals in this report are the grader's and are unaffected by what is withheld.

| EXAM              | DISCLOSURE CLASS | RULE APPLIED                                                                                                                                                                                                                                                                                                                         |
|-------------------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| image-object-1    | answer_key       | Applied consistently to 10 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published. |
| robot-video-1     | answer_key       | Applied consistently to 10 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published. |
| structured-data-1 | answer_key       | Applied consistently to 10 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published. |
| text-labeling-1   | answer_key       | Applied consistently to 10 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published. |

09 Attempt record 1 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-2d00c77f · record digest 648e95b3d5711091... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 42.1s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-2d00c77f                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:33:59 | Evaluation run started                               |
| 2026-09-06 03:34:41 | Evaluation run finished                              |
| 2026-09-06 03:34:41 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 40.9s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 5323bfa3-a60c-4768-8e5d-77011caa9f08

| Score | Items | Valid / invalid | Types                         |
|-------|-------|-----------------|-------------------------------|
| 1     | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |   |
|------------------|---|-------------------------|---|------------------------|---|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 1 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1 |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1 |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1 |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 0 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 1        | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T033359Z**

EXAM FINGERPRINT

**422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**648e95b3d571109148c78137d9202e454bec7f03164f20e1ac4d24ca823395e0**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**2d3c6cc47417289585f2a1673108bc17e48af7778016ad7e610a73f54028d419**

sha256 of fr-20260906-2d00c77f.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**50ca4f7ad369df48ce83ac24c06c4a36bcf8779a27909f9e28df195abaaa1fbb**

sha256 of fr-20260906-2d00c77f.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**63c925abbf5c820637fdc5b24c1a09c9620e244c26f7402b4d803b5a7408918**

sha256 of fr-20260906-2d00c77f.pdf.

RUN EVIDENCE BUNDLE

**2d00c77f2eafcddf5bb9f9811f4321e966b9c4db9d6b1b08a35191a8f5babe2e**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**c945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-6dc38b98. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 5323bfa3 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 40.9s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |



Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

2d00c77f2eafcddf5bb9f9811f4321e966b9c4db9d6b1b08a35191a8f5babe2e

RUN ID

live-20260906T033359Z

STARTED

2026-09-06T03:33:59.041472+00:00

FINISHED

2026-09-06T03:34:41.111670+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 0566648a2f0f5aa6f76d7c2f7099345b7225140726c138bf741a99a817b651a2 |
| rollouts/5323bfa3-a60c-4768-8e5d-77011caa9f08.json | 76bf5558574a5e2b6a0c8242576e36685a43f89ac2c03a21653880c922e3016a |
| spec                                               | a5839b46b35c23ec0db05ee9bed7df2b0f3f3cf3555b0c0ac6b6e097865c1e5c |
| summary                                            | 1e8bb7d67a4edeaf91e318b8a6e7e27a9f6abf4acac394f5a12b0eb5e06ca2c9 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

5323bfa3 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | read_file  | —                    | failed | 35    |
| 1    | list_files | .                    | ok     | 80    |
| 2    | read_file  | TASK.md              | ok     | 443   |
| 3    | read_file  | annotation-task.json | ok     | 3 093 |
| 4    | read_file  | assets/objects.png   | ok     | 397   |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (49 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

10 Attempt record 2 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-470625ed · record digest d1c1e3f5eaa431b0... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 44.6s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-470625ed                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:31:23 | Evaluation run started                               |
| 2026-09-06 03:32:07 | Evaluation run finished                              |
| 2026-09-06 03:32:07 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 43.5s     | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission e55d97b3-2be4-4529-9f24-9ce92ee21bf8

| Score  | Items | Valid / invalid | Types                         |
|--------|-------|-----------------|-------------------------------|
| 0.9878 | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |        |
|------------------|---|-------------------------|---|------------------------|--------|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 0.9756 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1      |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1      |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1      |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 1      |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.9756   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.9878

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033123Z

EXAM FINGERPRINT

422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**d1c1e3f5aaa431b0b2ae53bc  
cae13895768f3d52b8a7d98d  
db2f7f1ee6a53e74**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**1b78d7e4a415ab64a728d01f  
82a3b8840d3f26e4bbc4d6b7  
11d4df6557d77542**

sha256 of fr-20260906-470625ed.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**24539341d318543e74421f54  
73849765cb95b04c0e86bc62  
d5504b102220514e**

sha256 of fr-20260906-470625ed.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**89364390cc7a684aa6377f3e  
038c04a0b4b2d33156763a10  
bc00f12b3980617e**

sha256 of fr-20260906-470625ed.pdf.

RUN EVIDENCE BUNDLE

**470625edcc9bfadc81530685  
12aa7a787a9e722ad7e9b5d9  
3cee7ebc54b54424**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32f  
dd5fed0c4e781ab95eaaebfd  
e31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-aef86705`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| e55d97b3 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 43.5s   |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>470625edcc9bfadc8153068512aa7a787a9e722ad7e9b5d93cee7ebc54b54424</div></div> | <div><div>RUN ID</div><div>live-20260906T033123Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:31:23.385841+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:32:07.951623+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | a4fafaae45904f534a4ff2315127fe0c8e4c0a5676a63b90cdbac66c216d9450 |
| rollouts/e55d97b3-2be4-4529-9f24-9ce92ee21bf8.json | f74249ffe47571bee58666703b38de4c597ccc747fad0374820e337f2a5c43d5 |
| spec                                               | 83dae7ed85669e30d9ddf865bbc31d1155109b3c9cb42bab0f28bf094e3a3fb9 |
| summary                                            | fc8e6aecb39d82f1e18351fcfb31a8b4da5c589b0ca2be24490af95e0e80f2f4 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.



Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

e55d97b3 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (49 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

11 Attempt record 3 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-4a88a9bf · record digest 6947e974ea555f5a... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 43.5s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-4a88a9bf                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:28:53 | Evaluation run started                               |
| 2026-09-06 03:29:37 | Evaluation run finished                              |
| 2026-09-06 03:29:37 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 42.4s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 9c52dbfe-8f2a-40b5-b77f-c26f951a9f4c

| Score | Items | Valid / invalid | Types                         |
|-------|-------|-----------------|-------------------------------|
| 1     | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |   |
|------------------|---|-------------------------|---|------------------------|---|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 1 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1 |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1 |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1 |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 0 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.



Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 1        | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T032853Z**

EXAM FINGERPRINT

**422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**6947e974ea555f5a7d8f4664650e33c2badab9922358dbac8ee209231413d495**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**ad894127614870eed695ea5e31e7d0898bde1f4313c597316767bd50d032bd71**

sha256 of fr-20260906-4a88a9bf.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**1430ca18224935c7dd42b591256e9a4c13abfd2927596add8dc4d68fc4c7aabf**

sha256 of fr-20260906-4a88a9bf.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**236d8026b1a022c0565737634d09d548113af885aa4996abf9b1baf09c81c10e**

sha256 of fr-20260906-4a88a9bf.pdf.

RUN EVIDENCE BUNDLE

**4a88a9bf41b967135b1e1c2779f90b3d240b06f31ea26bb4774179aef1c51394**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-fd180ec1. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 9c52dbfe | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 42.4s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>4a88a9bf41b967135b1e1c2779f90b3d240b06f31ea26bb4774179aef1c51394</div></div> | <div><div>RUN ID</div><div>live-20260906T032853Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:28:53.799243+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:29:37.341255+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 5908f2ffb18d75bdb69679eeec4c1e6cadf29c9db27ff5a76dff885fd0b05e19 |
| rollouts/9c52dbfe-8f2a-40b5-b77f-c26f951a9f4c.json | 193c3bbfbf4f87d07cfad4596d6dd6d0caa0bf3bd672fa795e6ee68a798f2089 |
| spec                                               | c81ddd89da8c122d77dee1debd535662aa774c5b0fa811b160248d84d5143151 |
| summary                                            | f2c6756507abaffb8206223761dcbbe2453de312441d8448ea42c4bdd051608f |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

9c52dbfe · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (48 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.



12 Attempt record 4 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-62fb1341 · record digest 7a95898a5f2cd0f0... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.2s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-62fb1341                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:26:42 | Evaluation run started                               |
| 2026-09-06 03:27:29 | Evaluation run finished                              |
| 2026-09-06 03:27:29 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 45.8s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission aaff1bb0-dd17-45f4-9e5b-d47dc07237ae

| Score | Items | Valid / invalid | Types                         |
|-------|-------|-----------------|-------------------------------|
| 1     | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |   |
|------------------|---|-------------------------|---|------------------------|---|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 1 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1 |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1 |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1 |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 0 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 1        | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T032642Z**

EXAM FINGERPRINT

**422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**7a95898a5f2cd0f0b73c4e885cbbff22d2d90754e7602bb646a68df5f8f5a1e6**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**14a04a7582760fe0819c4872bf3c8a1cc07b4b4a45f334fb740d08cbb4ac17cb**

sha256 of `fr-20260906-62fb1341.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**046c4ac3478f0796a4e657eb6768c4847707f3ae5a9da14730bd78e7c304eefb**

sha256 of `fr-20260906-62fb1341.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**2a53b4f44c70010673f897f98af9a8795cfcad44734d6e4f45e8dda8cf7f02e6**

sha256 of `fr-20260906-62fb1341.pdf`.

RUN EVIDENCE BUNDLE

**62fb1341ce4fa28ba2c710132b84cfba742df77f008d6ea3877e233b3c5a5a17**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

This is the first recorded run on this exam. There is nothing to compare it against yet, and a single measurement is not a trend.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| aaff1bb0 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 45.8s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                          |                                                    |                                                                |                                                                 |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>62fb1341ce4fa28ba2c710132b84cfba742df77f008d6ea3877e233b3c5a5a17</div> | <div>RUN ID</div> <div>live-20260906T032642Z</div> | <div>STARTED</div> <div>2026-09-06T03:26:42.461545+00:00</div> | <div>FINISHED</div> <div>2026-09-06T03:27:29.639206+00:00</div> |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | c73265e660125e6ad5b1e5deb246050cf919aeedb525c577bd88e84465d66300 |
| rollouts/aaff1bb0-dd17-45f4-9e5b-d47dc07237ae.json | 3ad8b53071a2ff61a63f9dc78bc3fe6d2b02af701b172cc0fab64eb552c1ef8a |
| spec                                               | 4fac4e78aafc553bf40c4b0f1409779c9e8191a9f1255d8a15ea753250a3c93c |
| summary                                            | 6719ce07fccf7c71768b09ca6dab27ba11aa68778a1bf3bd9e6bee41fbe6819c |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

aaff1bb0 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (49 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

13 Attempt record 5 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-6dc38b98 · record digest a66b51f5da540d99... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 63.2s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-6dc38b98                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:32:55 | Evaluation run started                               |
| 2026-09-06 03:33:59 | Evaluation run finished                              |
| 2026-09-06 03:33:59 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 62.1s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission bae7aced-c973-4401-b7de-7493a2c82428

| Score | Items | Valid / invalid | Types                         |
|-------|-------|-----------------|-------------------------------|
| 1     | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |   |
|------------------|---|-------------------------|---|------------------------|---|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 1 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1 |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1 |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1 |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 0 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 1        | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.



Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASSEvidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.  
F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTEEvidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.  
F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033255Z

EXAM FINGERPRINT

422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**a66b51f5da540d99f9ecf6ff3602d204399934d919da8d7a9859684c63d583c9**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**feff90bb22def0b085a9e392dfff7e78048fe110df23e52c397d44fc2486ae95**

sha256 of `fr-20260906-6dc38b98.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**e9c3d51997548c6f4ec9b38c28ad98f2b96201affa32279c35aa9584b0775f1b**

sha256 of `fr-20260906-6dc38b98.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**925a796cd38fda8da6c2afc65c1a9d9e6010d836328497537842cc800c3d1e73**

sha256 of `fr-20260906-6dc38b98.pdf`.

RUN EVIDENCE BUNDLE

**6dc38b9845b950cb5dcfa9f325d80cb67a3e10a5ce975e708c58ddb0b1f06f2d**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfd e31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-ad6f18b9`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| bae7aced | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 62.1s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>6dc38b9845b950cb5dcfa9f3</div> <div>25d80cb67a3e10a5ce975e70</div> <div>8c58ddb0b1f06f2d</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T033255Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:32:55.777161</div> <div>+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:33:59.010609</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 7aa0e735b365146859943b0c76fd27589a4b926f93db1142220408094d804a8f |
| rollouts/bae7aced-c973-4401-b7de-7493a2c82428.json | cf5de9c0e04b5b3f4ee6052079718c3de72b9852cca602f89f621308bcbb945d |
| spec                                               | 8157bb7135b1fd921cee9dc27f385054528d53872a842f08b8ecf76323dd3302 |
| summary                                            | d1834222ce1f77acb54123bd109d97083adef65f5ba4b4741bf1bb7919fdbd9f |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

bae7aced · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (48 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

14 Attempt record 6 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-884851f3 · record digest 37a729b8483f10b4... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 52.4s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.



Certification evidence

**Why this exam is trustworthy.** Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-884851f3                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:29:37 | Evaluation run started                               |
| 2026-09-06 03:30:29 | Evaluation run finished                              |
| 2026-09-06 03:30:29 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 51.3s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 80de479d-7bf3-4f5d-8cad-c2fc6f1def90

| Score | Items | Valid / invalid | Types                         |
|-------|-------|-----------------|-------------------------------|
| 1     | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |   |
|------------------|---|-------------------------|---|------------------------|---|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 1 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1 |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1 |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1 |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 0 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 1        | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data



## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T032937Z

EXAM FINGERPRINT

422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**37a729b8483f10b49bdd80e7ba918abe57e5ddc3d6ba8be593d23d816e082d9c**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**a6e1797f29c5b9edfbf793037485edfc8b0d024d38d2b6dec4b714a5825bbaff**

sha256 of fr-20260906-884851f3.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**659f1fbf1a57d8d8543417890c76992245f816e0536363bdab1c9ff4165a837b**

sha256 of fr-20260906-884851f3.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**102da5965df789962e2f15bf3f059cec2f543da72bf4ba36aaf76c4a07b67ce4**

sha256 of fr-20260906-884851f3.pdf.

RUN EVIDENCE BUNDLE

**884851f36f9da73185b6621396c2477c60255d828546be2590e4810f16299d05**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-4a88a9bf`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 80de479d | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 51.3s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

884851f36f9da73185b6621396c2477c60255d828546be2590e4810f16299d05

RUN ID

live-20260906T032937Z

STARTED

2026-09-06T03:29:37.370304+00:00

FINISHED

2026-09-06T03:30:29.816091+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | ce8df413310422fbee9729a69451580c54b776522575f40358d9b189ef0707da |
| rollouts/80de479d-7bf3-4f5d-8cad-c2fc6f1def90.json | e7cc3618282c2d6f241d5d978f9c6d0ea18df1c9ac3aeba19c855cd594611310 |
| spec                                               | 3ca33d5d10d20e31f37a2ba96e524be73520bc055d2ccfe6f7fbd2b882695b7  |
| summary                                            | bdc06ef54ab0f101687bc0c0d5e6ccd946dd0a7f52f1401a665e5c5455152311 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

80de479d · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (49 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

15 Attempt record 7 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-ad6f18b9 · record digest ad3ad73c931070d5... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.8s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.



Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-ad6f18b9                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:32:07 | Evaluation run started                               |
| 2026-09-06 03:32:55 | Evaluation run finished                              |
| 2026-09-06 03:32:55 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 46.5s     | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission a470601f-1c98-499f-9b16-b5759fd1239b

| Score  | Items | Valid / invalid | Types                         |
|--------|-------|-----------------|-------------------------------|
| 0.9878 | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |        |
|------------------|---|-------------------------|---|------------------------|--------|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 0.9756 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1      |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1      |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1      |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 1      |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.9756   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.9878

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

**R-01** PROBLEM  
**Submitted annotations rejected by the independent grader.**  
  
AFFECTED LANES  
**cli/codex-gpt-5.5**  
  
WHAT WE WOULD LOOK AT FIRST  
Review the measured annotation metrics and validator error codes against the declared rubric.

**R-02** For anyone reading the percentages  
Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.



Verification

RUN ID

live-20260906T033207Z

EXAM FINGERPRINT

422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**ad3ad73c931070d584cb6d9c  
b4c645a177b416c9758efe1b  
62d1bfee32b293b6**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**dffbda1adb2d1182d71ed36c  
155f5474096029901eb67638  
8ed02d13466fc69a**

sha256 of fr-20260906-ad6f18b9.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**bf0dae7a691486b5b49091dc  
d6492bbb515a84e8c006e319  
3d6b0c8d39e55313**

sha256 of fr-20260906-ad6f18b9.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**690a4afb3c0c05e62b5b3c39  
a3ff1ae0d92534214a2cf475  
bba7c36aa5dd3970**

sha256 of fr-20260906-ad6f18b9.pdf.

RUN EVIDENCE BUNDLE

**ad6f18b99b25bac17364e7af  
8e60b2f4dea5b1acf1c1b3ef  
e112e13b1088c3ff**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32f  
dd5fed0c4e781ab95eaaebfd  
e31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-470625ed. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| a470601f | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 46.5s   |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST  
ad6f18b99b25bac17364e7af  
8e60b2f4dea5b1acf1c1b3ef  
e112e13b1088c3ff

RUN ID  
live-  
20260906T033207Z

STARTED  
2026-09-  
06T03:32:07.981657  
+00:00

FINISHED  
2026-09-  
06T03:32:55.745591  
+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 452d38c30b5e31215b25a4342f9fe04d0e7587b2015abbd717a4f8867c607361 |
| rollouts/a470601f-1c98-499f-9b16-b5759fd1239b.json | c9ca1ac48d8068c3065ebd1dc4717af4f3eb394007e5bc639459f22278501725 |
| spec                                               | c3ff978ed9d1442006d00396b9e887a623d1eb0c4183777e8372d5caa235d405 |
| summary                                            | d213668dd56ef59a605bc39da74700e11bae32e5e6def3022fa74dee479e2f94 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

a470601f · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (48 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

16 Attempt record 8 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-aef86705 · record digest cde2fc42622e4ef2... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 53.5s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-aef86705                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:30:29 | Evaluation run started                               |
| 2026-09-06 03:31:23 | Evaluation run finished                              |
| 2026-09-06 03:31:23 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.



Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 52.4s     | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 8754fd9e-a98a-4531-9c70-9d86fe51b379

| Score  | Items | Valid / invalid | Types                         |
|--------|-------|-----------------|-------------------------------|
| 0.9938 | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |        |
|------------------|---|-------------------------|---|------------------------|--------|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 0.9875 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1      |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1      |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1      |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 1      |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.9875   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.9938

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |  |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       |  | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |  |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        |  | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |  |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |  |                                                   |
| F-02                                                                                                                                                                                                                                       |  | ATTENTION                                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                                                                                                                                                                                                   |  |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        |  | Evidence: Appendix A · Appendix C                 |
| Submitted annotations rejected by the independent grader.                                                                                                                                                                                  |  |                                                   |
| F-02 · derived from the recorded run data                                                                                                                                                                                                  |  |                                                   |
| F-03                                                                                                                                                                                                                                       |  | NOTE                                              |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                                                                                                                        |  |                                                   |
| Severity: NOTE                                                                                                                                                                                                                             |  | Evidence: Appendix D                              |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.                                                                                            |  |                                                   |
| F-03 · derived from the recorded run data                                                                                                                                                                                                  |  |                                                   |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T033029Z**

EXAM FINGERPRINT

**422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**cde2fc42622e4ef27c1a9505f9a81637fcf0e48bb634045dc0dc9f33e4f87ba5**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**9f51a77bc7806247551a2085359e58af72f64884c66efe0024060bb1b49752c5**

sha256 of fr-20260906-aef86705.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**77c83c095dca3f876c30583f027281e814258dc5b1ca3974e8e6078b94d276ee**

sha256 of fr-20260906-aef86705.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**0a68d2d0ce1527b358d8f6d525a6eba9237f0b07fdb86d41e2b4aeacf1f86a2c**

sha256 of fr-20260906-aef86705.pdf.

RUN EVIDENCE BUNDLE

**aef86705c78b7068f0e7413bd9a5347f7d9b8cc228dd27551d593e964b2b67cf**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-884851f3. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.



Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 8754fd9e | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 52.4s   |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

aef86705c78b7068f0e7413bd9a5347f7d9b8cc228dd27551d593e964b2b67cf

RUN ID

live-20260906T033029Z

STARTED

2026-09-06T03:30:29.849136+00:00

FINISHED

2026-09-06T03:31:23.356794+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | eaae634727cd671d3f0531f97e2e8b1aa80a134ea5a0dc7001f907b0dd4d670f |
| rollouts/8754fd9e-a98a-4531-9c70-9d86fe51b379.json | 39ae9e543567a08709d106c2d65b263fa885f57ab5294e8b80304f587819ba69 |
| spec                                               | 4647031596c020fe042db608d5302a543a76ca8236275f74a8692288c5af6647 |
| summary                                            | 594a99756cd1f0cb347951033691527889df58e059d4bd7cf187113add04a46c |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

8754fd9e · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (48 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

17 Attempt record 9 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-af2506a7 · record digest b7b6915c0f62d8d5... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 43.1s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-af2506a7                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:27:29 | Evaluation run started                               |
| 2026-09-06 03:28:12 | Evaluation run finished                              |
| 2026-09-06 03:28:12 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 41.9s     | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 17671a13-64e4-42b8-9a90-98b7d789da54

| Score  | Items | Valid / invalid | Types                         |
|--------|-------|-----------------|-------------------------------|
| 0.9938 | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |        |
|------------------|---|-------------------------|---|------------------------|--------|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 0.9875 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1      |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1      |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1      |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 1      |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.9875   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.9938

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T032729Z**

EXAM FINGERPRINT

**422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf260b35a3ef**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**b7b6915c0f62d8d5942e7d4d50d86b3967f5e9e4ab77908f010e7f79ecba5497**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**e6f8a67d25df861142ea20342aa798b9cef27b8899346bb21428a3adfd04f3f5**

sha256 of `fr-20260906-af2506a7.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**3ddc0569aeb927c77c94dd07665c4890a1ef0ceaebeb73416dc8abd6279d2678**

sha256 of `fr-20260906-af2506a7.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**92401f3bf4b44130ac821d4d56f475f72022c006186d50bae8df1202fc7ac526**

sha256 of `fr-20260906-af2506a7.pdf`.

RUN EVIDENCE BUNDLE

**af2506a73576d7a7b551f4738446c41a4c226fff913a430f7966f807f3c843f3**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-62fb1341`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 17671a13 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 41.9s   |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |



Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

af2506a73576d7a7b551f4738446c41a4c226fff913a430f7966f807f3c843f3

RUN ID

live-20260906T032729Z

STARTED

2026-09-06T03:27:29.669686+00:00

FINISHED

2026-09-06T03:28:12.720700+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 0679c1f696a5b04eeee6e1dae19f2f5fae5edf009322b85f35b4d2bde29b7dec |
| rollouts/17671a13-64e4-42b8-9a90-98b7d789da54.json | b591ffb582654ece26ecd8d145a3e7cc47c3b41135f03c0e9bacd8d75f7de136 |
| spec                                               | 797317daecc54e9585e1d495b627d4205a1df7fcee764ad47543b3b46e791d72 |
| summary                                            | ba6ab309f2a66cdc4b9fe8168885d5ff5d24b5a7c190f414975443c92e3b6b14 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

17671a13 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (49 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

18 Attempt record 10 of 10 · image-object-1

vvdex.annotation.image-object-1 · record fr-20260906-fd180ec1 · record digest 6a5c91191a02c8ad... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.image-object-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

F-01 No evaluated lane passed the hidden tests in this run

F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)

F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 41.0s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.image-object-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                 |         |            |                 |
|---------------------------------|---------|------------|-----------------|
| EXAM                            | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.image-object-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-fd180ec1                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.image-object-1 · v0.1.0                              |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:28:12 | Evaluation run started                               |
| 2026-09-06 03:28:53 | Evaluation run finished                              |
| 2026-09-06 03:28:53 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 40.7s     | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: image. Rubric: 1.0.

A constructed image is reviewed for classification, object boxes, attributes and keypoints.

Submission 5dae9654-ee00-419b-879e-9b36c1bb2d9f

| Score  | Items | Valid / invalid | Types                         |
|--------|-------|-----------------|-------------------------------|
| 0.9878 | 4     | 4 / 0           | box, classification, keypoint |

Measured metrics

|                  |   |                         |   |                        |        |
|------------------|---|-------------------------|---|------------------------|--------|
| annotationCount  | 4 | boxF1                   | 1 | boxIoU                 | 0.9756 |
| boxPrecision     | 1 | boxRecall               | 1 | classificationAccuracy | 1      |
| classificationF1 | 1 | classificationPrecision | 1 | classificationRecall   | 1      |
| keypointF1       | 1 | keypointNormalizedError | 0 | keypointPrecision      | 1      |
| keypointRecall   | 1 | matchedCount            | 4 | mismatchCount          | 1      |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.9756   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| keypointPrecision       | recorded | N/A              | 1        | –      | –      |                  |
| keypointRecall          | recorded | N/A              | 1        | –      | –      |                  |
| keypointF1              | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.9878

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

**R-01** PROBLEM  
**Submitted annotations rejected by the independent grader.**

AFFECTED LANES  
**cli/codex-gpt-5.5**

WHAT WE WOULD LOOK AT FIRST  
Review the measured annotation metrics and validator error codes against the declared rubric.

**R-02 For anyone reading the percentages**

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

## Verification

RUN ID

live-20260906T032812Z

## EXAM FINGERPRINT

422753713aeb9f41abcc4135d637e3064cae1e770fc8d22ec9dacf2  
60b35a3ef

The contract digest this run was executed against.

## Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

## campaign-rendered chapter

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALD RECORD DIGEST

6a5c91191a02c8ada9bb6fdd  
380d95e2df6559de255866d0  
9d3af7b2cec1eb23

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

9a7b68aeefc7ae49f9dd9b27  
763fbb3e6ca36ac2f2dd8bd7  
5ba2115d1a0a9970

sha256 of fr-20260906-  
fd180ec1.json as written beside  
the standalone report.

SOURCE STANDALONE  
REPORT

```
7a0db0ac63145c3ef3a167ba
bebc6dae85b7889e2e9f25e6
214d96923424efca
```

sha256 of fr-20260906-  
fd180ec1.html — independently  
verifiable with its own digest.

SOURCE STANDALONE PDF

d82cdc24a229db7ba272dab0  
5be336da444b3ab4791b99b8  
1944d44cc766c396

sha256 of fr-20260906-  
fd180ec1.pdf.

## RUN EVIDENCE BUNDLE

fd180ec1f54ab91e7a07bba4  
a93a9a4374e49de91752fe40  
39e03d399e02e5f2

The digest the run's own bundle  
computed over itself.

CERTIFIED EVIDENCE SEAL

cf945b4838a266664300b32f  
dd5fed0c4e781ab95eaaebfd  
e31a0ce8ae2d7295

The exam's sealed evidence, established before this run.

## Measurement history

Previous run on this exam: fr-20260906-af2506a7. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 5dae9654 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 40.7s   |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                          |                                                    |                                                                |                                                                 |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>fd180ec1f54ab91e7a07bba4a93a9a4374e49de91752fe4039e03d399e02e5f2</div> | <div>RUN ID</div> <div>live-20260906T032812Z</div> | <div>STARTED</div> <div>2026-09-06T03:28:12.749616+00:00</div> | <div>FINISHED</div> <div>2026-09-06T03:28:53.774751+00:00</div> |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | e3e3b9ebd3d1fae0bddefba4dd0c947592c8a3810168ec32eeffdbabaf91c3f  |
| rollouts/5dae9654-ee00-419b-879e-9b36c1bb2d9f.json | 8e33b26c0340fa810d6f75f61b8bd02c91de15379d96595895676bd957c0ec8f |
| spec                                               | 11b741bbfcf022efce3c54d4e0134222f9753dd66604e7612a1f582bf6ecf6de |
| summary                                            | aafdb53ffb2114fb7244845e5a0c17596d76a6e05ac48c5d4f0191d9d7d6a37c |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.



Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

5dae9654 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 443   |
| 2    | read_file  | annotation-task.json | ok     | 3 093 |
| 3    | read_file  | assets/objects.png   | ok     | 397   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (48 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

19 Attempt record 1 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-08cf41e3 · record digest 4cb0e73aa85083f8... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 267.7s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-08cf41e3                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:31:55 | Evaluation run started                               |
| 2026-09-06 03:36:23 | Evaluation run finished                              |
| 2026-09-06 03:36:23 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 265.2s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 0506c45a-be16-435c-8218-3af3f0d65e92

| Score | Items | Valid / invalid | Types                               |
|-------|-------|-----------------|-------------------------------------|
| 0     | 6     | 0 / 6           | box, classification, event, segment |

Measured metrics

|         |              |
|---------|--------------|
| Metrics | Not measured |
|---------|--------------|

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

temporal\_order

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 9 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 1                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033155Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**4cb0e73aa85083f8e0ef903fd4d28f277a177b2774a66767e5635c84b75f8213**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**e22ff7c1c2a1cffddd14bd9fe30a3a9df294aa03cc26f63c96f1e8f195c767cc**

sha256 of fr-20260906-08cf41e3.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**f74dd440429ecc5bf7a004c9b647d29ccf8e5fe23805a46a01caf7f32a2efc00**

sha256 of fr-20260906-08cf41e3.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**4f79cfdabc8fd47e455fff183111edad9c627869c1a7e9b295d49c57ab2fc08**

sha256 of fr-20260906-08cf41e3.pdf.

RUN EVIDENCE BUNDLE

**08cf41e30d5d5edd9277064a87b533b306fcd7d0c08942a13d85e43eaff15a4d**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-5921101f. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 0506c45a | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 265.2s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

08cf41e30d5d5edd9277064a87b533b306fcd7d0c08942a13d85e43eaff15a4d

RUN ID

live-20260906T033155Z

STARTED

2026-09-06T03:31:55.771541+00:00

FINISHED

2026-09-06T03:36:23.510271+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 6ead2286a502c05d3fad527f85c1779cc6e1afbb24cc3e0e826604fd8c9af6bd |
| rollouts/0506c45a-be16-435c-8218-3af3f0d65e92.json | 06c00f2602a87d2312ca565be6be098e4e1ca1730d54265e1b219c66ae1e3732 |
| spec                                               | cebe99ff901457119977ee24852d095f5baf1a45f0a9eec274cc202fc1c3c771 |
| summary                                            | cce51c57bade0b91cbc37c26c5f11a975cfe7d6d7eb2a389c46f22ffc645e816 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

0506c45a · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | submission.json      | ok     | 74    |
| 4    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 5    | read_file  | assets/robot.webm    | ok     | 3 702 |
| 6    | write_file | submission.json      | ok     | 64    |
| 7    | run_tests  | —                    | ok     | 186   |
| 8    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (61 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

20 Attempt record 2 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-12f8c6af · record digest 5fdd4fe7918f8689... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 234.5s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-12f8c6af                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:41:00 | Evaluation run started                               |
| 2026-09-06 03:44:54 | Evaluation run finished                              |
| 2026-09-06 03:44:54 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 230.2s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 152384db-2848-40ec-b854-21f8ce8bd511

| Score | Items | Valid / invalid | Types                      |
|-------|-------|-----------------|----------------------------|
| 0     | 6     | 0 / 6           | box, classification, event |

Measured metrics

|         |              |
|---------|--------------|
| Metrics | Not measured |
|---------|--------------|

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

missing\_type

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |



Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034100Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

5fdd4fe7918f8689e1ad975912f7241d661247fdccbc3b30aca8b96499b7a427

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

68efbba7425a5a2674275b4003e9418721987c9bf15edf33b59271849ccaa229

sha256 of fr-20260906-12f8c6af.json as written beside the standalone report.

SOURCE STANDALONE REPORT

239a6d157e740920d336daa9da7d830f921f8ccebcbf9dffa59642265d41aefd

sha256 of fr-20260906-12f8c6af.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

d64c01be303fce2674b1a7ccd9e549367bdbb656e9b74d7a9b8de76e8632edda

sha256 of fr-20260906-12f8c6af.pdf.

RUN EVIDENCE BUNDLE

12f8c6afa64cbb17fed0bef95a4b8ad7c9e47965549be915fdfd71299d85bddf

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-55c7c9bc. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 152384db | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 230.2s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>12f8c6afa64cbb17fed0bef95a4b8ad7c9e47965549be915fdfd71299d85bddf</div></div> | <div><div>RUN ID</div><div>live-20260906T034100Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:41:00.354844+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:44:54.908592+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | e40872c9d457b17b68e6938a7f5576941df860fdcdea2aca982c380b15530077 |
| rollouts/152384db-2848-40ec-b854-21f8ce8bd511.json | 340f664988e7ad0ca9d6b4f399d48ba43b08e56503719cff3c484d276cc401e9 |
| spec                                               | 740990ea9d9ce7352d8285af41b7036fc0a2f6b8813ee75ae57120106e3f2522 |
| summary                                            | 0ea8fb5f3224c3ed636a849b9610d8cf3e50c7dc59af71118f1f955aa93ccc57 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

152384db · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | assets/robot.webm    | ok     | 3 342 |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

21 Attempt record 3 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-55c7c9bc · record digest c50dca52eb4e3d73... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 276.8s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.



Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-55c7c9bc                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:36:23 | Evaluation run started                               |
| 2026-09-06 03:41:00 | Evaluation run finished                              |
| 2026-09-06 03:41:00 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 274.5s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 39656c04-9d24-45d4-9e13-2cb7519bc924

| Score  | Items | Valid / invalid | Types                               |
|--------|-------|-----------------|-------------------------------------|
| 0.4907 | 6     | 6 / 0           | box, classification, event, segment |

Measured metrics

|                    |        |                                    |     |                        |        |
|--------------------|--------|------------------------------------|-----|------------------------|--------|
| annotationCount    | 6      | boxF1                              | 0.5 | boxIoU                 | 0.5186 |
| boxPrecision       | 0.5    | boxRecall                          | 0.5 | classificationAccuracy | 1      |
| classificationF1   | 1      | classificationPrecision            | 1   | classificationRecall   | 1      |
| eventF1            | 0      | eventPrecision                     | 0   | eventRecall            | 0      |
| eventTemporalIoU   | 0.0286 | matchedCount                       | 5   | mismatchCount          | 5      |
| segmentF1          | 0.6667 | segmentPrecision                   | 0.5 | segmentRecall          | 1      |
| segmentTemporalIoU | 0.425  | Track consistency on matched boxes | 1   |                        |        |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 0.5      | –      | –      |                  |
| boxRecall               | recorded | N/A              | 0.5      | –      | –      |                  |
| boxF1                   | recorded | N/A              | 0.5      | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.5186   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| eventPrecision          | recorded | N/A              | 0        | –      | –      |                  |
| eventRecall             | recorded | N/A              | 0        | –      | –      |                  |
| eventF1                 | recorded | N/A              | 0        | –      | –      |                  |
| eventTemporalIoU        | recorded | N/A              | 0.0286   | –      | –      |                  |
| segmentPrecision        | recorded | N/A              | 0.5      | –      | –      |                  |
| segmentRecall           | recorded | N/A              | 1        | –      | –      |                  |
| segmentF1               | recorded | N/A              | 0.6667   | –      | –      |                  |
| segmentTemporalIoU      | recorded | N/A              | 0.425    | –      | –      |                  |
| trackingConsistency     | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.4907

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 9 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.



Verification

RUN ID

live-20260906T033623Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**c50dca52eb4e3d7351d38238b8f67be3139846d39eae70d6fb9f3e7edef1d8da**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**d2614f9716e0a056aa4d4818f0310e262c747e81e7ccad46406012d7e9c6e4b6**

sha256 of fr-20260906-55c7c9bc.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**4032259e05ccbd346661038d1ed62a8b74dd96273e8d9473da11cd65f21fa560**

sha256 of fr-20260906-55c7c9bc.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**4b271a8fecf56f374ed81a4c87d6bbf843f81b54df02398b2048a2ee99b01caa**

sha256 of fr-20260906-55c7c9bc.pdf.

RUN EVIDENCE BUNDLE

**55c7c9bc10b6d87dc539e125b62d8b525f7b9fe9c0c70b79db8797502030788b**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-08cf41e3. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 39656c04 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 274.5s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>55c7c9bc10b6d87dc539e125</div> <div>b62d8b525f7b9fe9c0c70b79</div> <div>db8797502030788b</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T033623Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:36:23.55066</div> <div>2+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:41:00.312413</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 907da3228e3dbedc6addd26253b1c2010847aacd6036cf04e312b28b7da45957 |
| rollouts/39656c04-9d24-45d4-9e13-2cb7519bc924.json | 663ff4ae427c1614ee2c9ea0f9ee757a79cba676d3c2e5c2dd97245c8057d472 |
| spec                                               | 77b7232d62f78d1f9f3ef3b375774100ea767c6bf03c4090ad7e5af886b86aa8 |
| summary                                            | 0d56af262bb48963172b1fbd664a5cada933fdb256cd3fb68d46d79bae55445a |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

39656c04 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | read_file  | assets/robot.webm    | ok     | 3 702 |
| 6    | write_file | submission.json      | ok     | 64    |
| 7    | run_tests  | —                    | ok     | 186   |
| 8    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (61 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

22 Attempt record 4 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-57451e50 · record digest bde0ebe1b4f14b79... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 256.3s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-57451e50                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 04:00:54 | Evaluation run started                               |
| 2026-09-06 04:05:10 | Evaluation run finished                              |
| 2026-09-06 04:05:10 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.



Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 251.8s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 0ad9e41b-6ba3-4f96-8b55-39dad8023c13

| Score  | Items | Valid / invalid | Types                               |
|--------|-------|-----------------|-------------------------------------|
| 0.4583 | 6     | 6 / 0           | box, classification, event, segment |

Measured metrics

|                    |        |                                    |     |                        |       |
|--------------------|--------|------------------------------------|-----|------------------------|-------|
| annotationCount    | 6      | boxF1                              | 0.5 | boxIoU                 | 0.461 |
| boxPrecision       | 0.5    | boxRecall                          | 0.5 | classificationAccuracy | 1     |
| classificationF1   | 1      | classificationPrecision            | 1   | classificationRecall   | 1     |
| eventF1            | 0      | eventPrecision                     | 0   | eventRecall            | 0     |
| eventTemporalIoU   | 0.0238 | matchedCount                       | 5   | mismatchCount          | 5     |
| segmentF1          | 0.6667 | segmentPrecision                   | 0.5 | segmentRecall          | 1     |
| segmentTemporalIoU | 0.39   | Track consistency on matched boxes | 1   |                        |       |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 0.5      | –      | –      |                  |
| boxRecall               | recorded | N/A              | 0.5      | –      | –      |                  |
| boxF1                   | recorded | N/A              | 0.5      | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.461    | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| eventPrecision          | recorded | N/A              | 0        | –      | –      |                  |
| eventRecall             | recorded | N/A              | 0        | –      | –      |                  |
| eventF1                 | recorded | N/A              | 0        | –      | –      |                  |
| eventTemporalIoU        | recorded | N/A              | 0.0238   | –      | –      |                  |
| segmentPrecision        | recorded | N/A              | 0.5      | –      | –      |                  |
| segmentRecall           | recorded | N/A              | 1        | –      | –      |                  |
| segmentF1               | recorded | N/A              | 0.6667   | –      | –      |                  |
| segmentTemporalIoU      | recorded | N/A              | 0.39     | –      | –      |                  |
| trackingConsistency     | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.4583

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T040054Z**

EXAM FINGERPRINT

**c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**bde0ebe1b4f14b7989a966b5a11157dc19cb88391c5937324b23d81d9da1a61b**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**336ae02ef9f69419f8d6afb70dceba7625b832c2f9c86f28290bed67a764fad7**

sha256 of fr-20260906-57451e50.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**383833b7645ef90a6fc557fc c85505ae9dc5dc45382cc0454fc1d0d11a9448d3**

sha256 of fr-20260906-57451e50.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**a8e1f8cea2d77ce7246b0ba6f4cebe64de265742be73a91690c2c232e7b67556**

sha256 of fr-20260906-57451e50.pdf.

RUN EVIDENCE BUNDLE

**57451e5016d5832be0b5fd6b12d5a2e98c316b63e2d04944e676b6aba1f57551**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-7b7f8fb2. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.



Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 0ad9e41b | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 251.8s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                          |                                                    |                                                                |                                                                 |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>57451e5016d5832be0b5fd6b12d5a2e98c316b63e2d04944e676b6aba1f57551</div> | <div>RUN ID</div> <div>live-20260906T040054Z</div> | <div>STARTED</div> <div>2026-09-06T04:00:54.539879+00:00</div> | <div>FINISHED</div> <div>2026-09-06T04:05:10.868337+00:00</div> |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | d6ecae527dfc56b5a204a6141e331fbaad4516c743bd9db6bef666ae6c85e94c |
| rollouts/0ad9e41b-6ba3-4f96-8b55-39dad8023c13.json | 87cd943066f3ad01d85a2fe6e444bdd4869854adfe6fa957c971a81b738ac6e3 |
| spec                                               | bc53103201670d99b9638f4157d9d2e4414613512ab76bf1efc321d550f187c7 |
| summary                                            | 9257ef55e0353e6717b8f736db28bf022a401b03b25787413bab08ee0109403d |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

0ad9e41b · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | assets/robot.webm    | ok     | 3 342 |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

23 Attempt record 5 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-5921101f · record digest cafc99313b275331... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 187.4s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-5921101f                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:28:48 | Evaluation run started                               |
| 2026-09-06 03:31:55 | Evaluation run finished                              |
| 2026-09-06 03:31:55 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                                   | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|-----------------------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| <b>GPT-5.5 via Codex CLI</b><br><b>(subscription, image input)</b><br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 185.2s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 631337e7-8b67-446e-a203-e32e2f7b9cc3

| Score | Items | Valid / invalid | Types                               |
|-------|-------|-----------------|-------------------------------------|
| 0.581 | 6     | 6 / 0           | box, classification, event, segment |

Measured metrics

|                    |        |                                    |   |                        |       |
|--------------------|--------|------------------------------------|---|------------------------|-------|
| annotationCount    | 6      | boxF1                              | 1 | boxIoU                 | 0.601 |
| boxPrecision       | 1      | boxRecall                          | 1 | classificationAccuracy | 1     |
| classificationF1   | 1      | classificationPrecision            | 1 | classificationRecall   | 1     |
| eventF1            | 0      | eventPrecision                     | 0 | eventRecall            | 0     |
| eventTemporalIoU   | 0.25   | matchedCount                       | 5 | mismatchCount          | 5     |
| segmentF1          | 1      | segmentPrecision                   | 1 | segmentRecall          | 1     |
| segmentTemporalIoU | 0.7839 | Track consistency on matched boxes | 1 |                        |       |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.601    | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| eventPrecision          | recorded | N/A              | 0        | –      | –      |                  |
| eventRecall             | recorded | N/A              | 0        | –      | –      |                  |
| eventF1                 | recorded | N/A              | 0        | –      | –      |                  |
| eventTemporalIoU        | recorded | N/A              | 0.25     | –      | –      |                  |
| segmentPrecision        | recorded | N/A              | 1        | –      | –      |                  |
| segmentRecall           | recorded | N/A              | 1        | –      | –      |                  |
| segmentF1               | recorded | N/A              | 1        | –      | –      |                  |
| segmentTemporalIoU      | recorded | N/A              | 0.7839   | –      | –      |                  |
| trackingConsistency     | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.581

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T032848Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**cafc99313b275331a9ae86c70be3f042b7ab7b8d0b442a9d5100f51c3ba81e93**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**f94d8ad16ca13f05b952a30acdd19fa1d3bb7b6083824f5003c325edfbd7383d**

sha256 of fr-20260906-5921101f.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**02fa9fab4712e50b055681875441dc317c7992eb515ef4d7a59a6c779bee4b28**

sha256 of fr-20260906-5921101f.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**1747e3c2a29ab6bc418973e92276052f8f9a9374c04c5247483e5609171c17a8**

sha256 of fr-20260906-5921101f.pdf.

RUN EVIDENCE BUNDLE

**5921101f770d0271feaccbbc03f881e53de1f9480afd55b15fbf894673184caa**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-ee230324. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 631337e7 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 185.2s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |



Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

5921101f770d0271feaccbbc

03f881e53de1f9480afd55b1

5fbf894673184caa

RUN ID

live-

20260906T032848Z

STARTED

2026-09-

06T03:28:48.316941

+00:00

FINISHED

2026-09-

06T03:31:55.732318

+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 89f7637cec8cd86d985cc3aa5be0cd4a4c5810d5498afb21e25793a7a719703a |
| rollouts/631337e7-8b67-446e-a203-e32e2f7b9cc3.json | 377d9f8b6ff0d336e131a6d845a7d6cf9848acc82c7b2fc8b98f2292fa1687a6 |
| spec                                               | f054d76bb343f5196c88cd6fba869e8bbcf820b1074b913c6b53ee78ae77fc2c |
| summary                                            | 9bf91a52d89a7358261494fe8fde5084719aebec58d05f0e23d919295aba438b |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

631337e7 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | submission.json      | ok     | 74    |
| 4    | read_file  | assets/robot.webm    | ok     | 3 342 |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

24 Attempt record 6 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-7b7f8fb2 · record digest 1879a1fa6919d32c... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01No evaluated lane passed the hidden tests in this run
- F-02submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 286.4s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

**Why this exam is trustworthy.** Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-7b7f8fb2                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:56:08 | Evaluation run started                               |
| 2026-09-06 04:00:54 | Evaluation run finished                              |
| 2026-09-06 04:00:54 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                                   | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|-----------------------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| <b>GPT-5.5 via Codex CLI</b><br><b>(subscription, image input)</b><br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 283.2s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission ba135765-48e1-4ab1-ab9d-973a6fcfcdd6

| Score | Items | Valid / invalid | Types                               |
|-------|-------|-----------------|-------------------------------------|
| 0     | 6     | 0 / 6           | box, classification, event, segment |

Measured metrics

|         |              |
|---------|--------------|
| Metrics | Not measured |
|---------|--------------|

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

temporal\_order

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 9 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T035608Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

campaign-rendered chapter

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

1879a1fa6919d32cd484a4e39530ec262f54bdc00e7b0b1015ed4d2032245145

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

3f61012ac919be872136fa8abbafc7a56fa249fc6137370619aa377f57be1f88

sha256 of fr-20260906-7b7f8fb2.json as written beside the standalone report.

SOURCE STANDALONE REPORT

ccb3e8d90c5226e56d8522ed1b1c6738f0e9fcb643f9ec4438db234214d10d4a

sha256 of fr-20260906-7b7f8fb2.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

a1c9a114fd4dc9f08bb597a51a0afef93f97f81500c6217d1167312128ea544b

sha256 of fr-20260906-7b7f8fb2.pdf.

RUN EVIDENCE BUNDLE

7b7f8fb22fecdbc1198ab9c6ba89e7d457a5601aedac1d15cd8c4c284d5b7af6

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-7f86cd6e. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| ba135765 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 283.2s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>7b7f8fb22fecdbc1198ab9c6ba89e7d457a5601aedac1d15cd8c4c284d5b7af6</div></div> | <div><div>RUN ID</div><div>live-20260906T035608Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:56:08.099510+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T04:00:54.506742+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 264ffcdbb813e0987a5d3453a4039d41bbb87afac0d407688816f1a9df89023  |
| rollouts/ba135765-48e1-4ab1-ab9d-973a6fcfcdd6.json | 882472b3628b2c3deadf1207c806fa87ac379bb83c452964f2395ccbc1c5e777 |
| spec                                               | c16ee74b6edae6328198c57a5ce1ffa9a0c20f19f38858f4eeeeeece2d002a28 |
| summary                                            | b0ecca619b818d4451180ea8435fc6ecb611cd1c8275038ca9f0deb84a9d2c1e |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

ba135765 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | read_file  | assets/robot.webm    | ok     | 3 702 |
| 6    | write_file | submission.json      | ok     | 64    |
| 7    | run_tests  | —                    | ok     | 186   |
| 8    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (61 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.



25 Attempt record 7 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-7f86cd6e · record digest 6637ce529278ae9c... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 123.6s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-7f86cd6e                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:54:04 | Evaluation run started                               |
| 2026-09-06 03:56:08 | Evaluation run finished                              |
| 2026-09-06 03:56:08 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 121.6s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 0e45e6f6-355a-4920-a121-5a0879b720c7

| Score  | Items | Valid / invalid | Types                               |
|--------|-------|-----------------|-------------------------------------|
| 0.6266 | 6     | 6 / 0           | box, classification, event, segment |

Measured metrics

|                    |        |                                    |     |                        |        |
|--------------------|--------|------------------------------------|-----|------------------------|--------|
| annotationCount    | 6      | boxF1                              | 1   | boxIoU                 | 0.6314 |
| boxPrecision       | 1      | boxRecall                          | 1   | classificationAccuracy | 1      |
| classificationF1   | 1      | classificationPrecision            | 1   | classificationRecall   | 1      |
| eventF1            | 0.5    | eventPrecision                     | 0.5 | eventRecall            | 0.5    |
| eventTemporalIoU   | 0.3385 | matchedCount                       | 6   | mismatchCount          | 5      |
| segmentF1          | 1      | segmentPrecision                   | 1   | segmentRecall          | 1      |
| segmentTemporalIoU | 0.82   | Track consistency on matched boxes | 1   |                        |        |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.6314   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| eventPrecision          | recorded | N/A              | 0.5      | –      | –      |                  |
| eventRecall             | recorded | N/A              | 0.5      | –      | –      |                  |
| eventF1                 | recorded | N/A              | 0.5      | –      | –      |                  |
| eventTemporalIoU        | recorded | N/A              | 0.3385   | –      | –      |                  |
| segmentPrecision        | recorded | N/A              | 1        | –      | –      |                  |
| segmentRecall           | recorded | N/A              | 1        | –      | –      |                  |
| segmentF1               | recorded | N/A              | 1        | –      | –      |                  |
| segmentTemporalIoU      | recorded | N/A              | 0.82     | –      | –      |                  |
| trackingConsistency     | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.6266

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 10 / 24    | 0                | 0 / 0    | 0                | 0                     | 0        | 1                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T035404Z**

EXAM FINGERPRINT

**c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**6637ce529278ae9c381a7407303cca0aa6aadfa60671a0efdedf2ba76e576f18**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**18578f1998955d663eef348cba5295932c19ea2a39489feb0151e4fbca902458**

sha256 of `fr-20260906-7f86cd6e.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**d4742139ee280853091b4ff84d5c1b99b5822d9d9be96eda37a37df5edfc9bd1**

sha256 of `fr-20260906-7f86cd6e.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**52daab990f3cf3b1937d8c327dc36ea1e79b62cb2ed6d59f31cbc6efbcf5e340**

sha256 of `fr-20260906-7f86cd6e.pdf`.

RUN EVIDENCE BUNDLE

**7f86cd6ec409647b2312dd075f22b9f36cf615dabf610899a411855a3b706c75**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-c5386852`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 0e45e6f6 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 121.6s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>7f86cd6ec409647b2312dd07</div> <div>5f22b9f36cf615dabf610899</div> <div>a411855a3b706c75</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T035404Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:54:04.44206</div> <div>6+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:56:08.06598</div> <div>8+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | fbf42907d1e92a63a7576780bd4345b806773b921a45a0ec8056a1000f00f0d2 |
| rollouts/0e45e6f6-355a-4920-a121-5a0879b720c7.json | b9528827b07edf544cf6e8541d276c09c7e173f5b2607871b4c1b631e7e70d13 |
| spec                                               | c42ce090047f2bba44a337ffe8b11d79b5ee80736cc3c2264bff30cc87ead363 |
| summary                                            | 40c94d28c15d9b979d7028c4976460c9eff266647a1f1e5bc64e84e09bd5670f |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

0e45e6f6 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | read_file  | —                    | failed | 35    |
| 1    | list_files | .                    | ok     | 79    |
| 2    | read_file  | TASK.md              | ok     | 732   |
| 3    | read_file  | annotation-task.json | ok     | 4 329 |
| 4    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 5    | read_file  | assets/robot.webm    | ok     | 822   |
| 6    | read_file  | submission.json      | ok     | 74    |
| 7    | write_file | submission.json      | ok     | 64    |
| 8    | run_tests  | —                    | ok     | 186   |
| 9    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

26 Attempt record 8 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-c5386852 · record digest 014e24edff39b7c3... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01No evaluated lane passed the hidden tests in this run
- F-02submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 293.9s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-c5386852                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:49:10 | Evaluation run started                               |
| 2026-09-06 03:54:04 | Evaluation run finished                              |
| 2026-09-06 03:54:04 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                  |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                        |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file, write_file, edit_file, run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 289.9s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 14bdfb89-a386-4a47-8e85-9827b56428d0

| Score | Items | Valid / invalid | Types                      |
|-------|-------|-----------------|----------------------------|
| 0     | 6     | 0 / 6           | box, classification, event |

Measured metrics

|         |              |
|---------|--------------|
| Metrics | Not measured |
|---------|--------------|

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

missing\_type

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 10 / 24    | 0                | 0 / 0    | 0                | 0                     | 0        | 1                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |



Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034910Z

EXAM FINGERPRINT

c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**014e24edff39b7c34708f0bc5e8f1c826ba7e8ec940bf705cbb0f96f22d7a832**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**d4fa7f49699a6bca498d49c6fe4bee1f3ad5cab9d9149c31b5cea98f6cafb171**

sha256 of `fr-20260906-c5386852.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**c356fa31896ad36ec00963c1be17456cbfab15d8658f17f76939c0ccb79ea610**

sha256 of `fr-20260906-c5386852.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**1be13eb3996e4f76e0e27b26ee0a59a4f434ac6d844518a9e79544cfe8e72ca7**

sha256 of `fr-20260906-c5386852.pdf`.

RUN EVIDENCE BUNDLE

**c5386852f0e67e7b11edad9b8320da6f4ffb043f308d69aca37d777c66ed1bad**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-ea40e0a7`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 14bdfb89 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 289.9s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

c5386852f0e67e7b11edad9b8320da6f4ffb043f308d69aca37d777c66ed1bad

RUN ID

live-20260906T034910Z

STARTED

2026-09-06T03:49:10.546517+00:00

FINISHED

2026-09-06T03:54:04.411735+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | c87ba4b51fddbc3082194fd6815385ae58bd0051213b3b7b1dbb01d1811f1e92 |
| rollouts/14bdfb89-a386-4a47-8e85-9827b56428d0.json | 579feda8ad927fccf7d3f2a2306c5d922336dfbbdd85d6aa772161d40328b6cb |
| spec                                               | 3d195039341613649270f4d74313cbc57713fc5eabccd88eef546ae99fa4ba7d |
| summary                                            | b8ae91ec01783bc9cdee86ab545e47fdfa727b266f131104c4b346c761689b78 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

14bdfb89 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | read_file  | —                    | failed | 35    |
| 1    | list_files | .                    | ok     | 79    |
| 2    | read_file  | TASK.md              | ok     | 732   |
| 3    | read_file  | annotation-task.json | ok     | 4 329 |
| 4    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 5    | read_file  | assets/robot.webm    | ok     | 3 342 |
| 6    | read_file  | submission.json      | ok     | 74    |
| 7    | write_file | submission.json      | ok     | 64    |
| 8    | run_tests  | —                    | ok     | 186   |
| 9    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

27 Attempt record 9 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-ea40e0a7 · record digest 6b7b5285b5fc6dd7... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>0</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 255.5s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.



Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-ea40e0a7                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:44:54 | Evaluation run started                               |
| 2026-09-06 03:49:10 | Evaluation run finished                              |
| 2026-09-06 03:49:10 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                                   | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|-----------------------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| <b>GPT-5.5 via Codex CLI</b><br><b>(subscription, image input)</b><br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 251.6s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 5228869f-3331-424a-be35-6d623497241a

| Score | Items | Valid / invalid | Types                      |
|-------|-------|-----------------|----------------------------|
| 0     | 6     | 0 / 6           | box, classification, event |

Measured metrics

|         |              |
|---------|--------------|
| Metrics | Not measured |
|---------|--------------|

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

missing\_type

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 9 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034454Z**

EXAM FINGERPRINT

**c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**6b7b5285b5fc6dd7c66338dc973344ad8a80f87314711b7ae9e95f783f8ad1b5**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**bed482d18f5123d104ab20e66a4743cf3f0691171737a3836d6f4fa1d734b5ff**

sha256 of `fr-20260906-ea40e0a7.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**876b704b31757befc08e60c6f47413e634b5dd13a6ce8d3aa9f03bbab59ec451**

sha256 of `fr-20260906-ea40e0a7.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**d33cbf84ef1ba25c9b4894dd8e1398d7728f605d2c7f85e21530ccba4e6ae2f6**

sha256 of `fr-20260906-ea40e0a7.pdf`.

RUN EVIDENCE BUNDLE

**ea40e0a76f232f5f91b5e0129614ff6d17630a65061913138b206a060c9bec51**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debf0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-12f8c6af`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.



Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 5228869f | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 251.6s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

ea40e0a76f232f5f91b5e0129614ff6d17630a65061913138b206a060c9bec51

RUN ID

live-20260906T034454Z

STARTED

2026-09-06T03:44:54.972101+00:00

FINISHED

2026-09-06T03:49:10.509984+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | fc19a049717f2cf24280bee5c8b7dc8fef578efaa5f0d418571f36d2c48818ca |
| rollouts/5228869f-3331-424a-be35-6d623497241a.json | b7e2514fc4d6b8418fae81e080585058bdf65e038b553b0acc139f9d1058b8f7 |
| spec                                               | db58aacde5c1d105ed590f6ca736bdc0cb2038e50859f6e1759e46d20126b984 |
| summary                                            | a27106a4e73e5322c606d4ea61d607a524c0c3dade2480076dd71f75c4f246df |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

5228869f · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | read_file  | —                    | failed | 35    |
| 1    | list_files | .                    | ok     | 79    |
| 2    | read_file  | TASK.md              | ok     | 732   |
| 3    | read_file  | annotation-task.json | ok     | 4 329 |
| 4    | read_file  | assets/robot.webm    | ok     | 3 342 |
| 5    | read_file  | submission.json      | ok     | 74    |
| 6    | write_file | submission.json      | ok     | 64    |
| 7    | run_tests  | —                    | ok     | 186   |
| 8    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (61 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

28 Attempt record 10 of 10 · robot-video-1

vvdex.annotation.robot-video-1 · record fr-20260906-ee230324 · record digest f041ab53c0abd3fb... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

0

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

0 of 1

graded rollouts passed the independent hidden tests and were certified

0 of 1 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was submitted\_failed — submitted annotations rejected by the independent grader.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 submitted\_failed × 1 (cli/codex-gpt-5.5)
- F-03 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 125.8s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                |         |            |                 |
|--------------------------------|---------|------------|-----------------|
| EXAM                           | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.robot-video-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1

EXAM FINGERPRINT

c3b4773c375c10ad989357bb...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-ee230324                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.robot-video-1 · v0.1.0                               |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:26:42 | Evaluation run started                               |
| 2026-09-06 03:28:48 | Evaluation run finished                              |
| 2026-09-06 03:28:48 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                  |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                        |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file, write_file, edit_file, run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 0 / 1  | 1           | 0           | 1 / 1     | 123.6s    | Graded                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 0      | 0.0% | [0%, 79%]    | 0     | 0.0%   |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 0 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Graded                | Graded                  | 0 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: video. Rubric: 1.0.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission 9e118309-ded3-4dc1-a44b-dfa232943eda

| Score | Items | Valid / invalid | Types                               |
|-------|-------|-----------------|-------------------------------------|
| 0.644 | 6     | 6 / 0           | box, classification, event, segment |

Measured metrics

|                    |        |                                    |     |                        |        |
|--------------------|--------|------------------------------------|-----|------------------------|--------|
| annotationCount    | 6      | boxF1                              | 1   | boxIoU                 | 0.6755 |
| boxPrecision       | 1      | boxRecall                          | 1   | classificationAccuracy | 1      |
| classificationF1   | 1      | classificationPrecision            | 1   | classificationRecall   | 1      |
| eventF1            | 0.5    | eventPrecision                     | 0.5 | eventRecall            | 0.5    |
| eventTemporalIoU   | 0.3664 | matchedCount                       | 6   | mismatchCount          | 5      |
| segmentF1          | 1      | segmentPrecision                   | 1   | segmentRecall          | 1      |
| segmentTemporalIoU | 0.78   | Track consistency on matched boxes | 1   |                        |        |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| boxPrecision            | recorded | N/A              | 1        | –      | –      |                  |
| boxRecall               | recorded | N/A              | 1        | –      | –      |                  |
| boxF1                   | recorded | N/A              | 1        | –      | –      |                  |
| boxIoU                  | recorded | N/A              | 0.6755   | –      | –      |                  |
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| eventPrecision          | recorded | N/A              | 0.5      | –      | –      |                  |
| eventRecall             | recorded | N/A              | 0.5      | –      | –      |                  |
| eventF1                 | recorded | N/A              | 0.5      | –      | –      |                  |
| eventTemporalIoU        | recorded | N/A              | 0.3664   | –      | –      |                  |
| segmentPrecision        | recorded | N/A              | 1        | –      | –      |                  |
| segmentRecall           | recorded | N/A              | 1        | –      | –      |                  |
| segmentF1               | recorded | N/A              | 1        | –      | –      |                  |
| segmentTemporalIoU      | recorded | N/A              | 0.78     | –      | –      |                  |
| trackingConsistency     | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

0.644

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 9 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 1                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

|                                                                                                                                                                                                                                            |                                                   |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------|
| F-01                                                                                                                                                                                                                                       | ATTENTION                                         |
| No evaluated lane passed the hidden tests in this run                                                                                                                                                                                      |                                                   |
| Severity: ATTENTION                                                                                                                                                                                                                        | Evidence: Model outcomes · Certification evidence |
| Every rollout either stopped before submitting or submitted an annotation submission the independent grader rejected. The exam's own certification (gold solution passes, empty submission fails) establishes the task itself is winnable. |                                                   |
| F-01 · derived from the recorded run data                                                                                                                                                                                                  |                                                   |

|                                                           |                                   |
|-----------------------------------------------------------|-----------------------------------|
| F-02                                                      | ATTENTION                         |
| submitted_failed × 1 (cli/codex-gpt-5.5)                  |                                   |
| Severity: ATTENTION                                       | Evidence: Appendix A · Appendix C |
| Submitted annotations rejected by the independent grader. |                                   |
| F-02 · derived from the recorded run data                 |                                   |

|                                                                                                                                                 |                      |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|
| F-03                                                                                                                                            | NOTE                 |
| Statistical floor — 1 rollout(s) is below the 10-rollout rate floor                                                                             |                      |
| Severity: NOTE                                                                                                                                  | Evidence: Appendix D |
| The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run. |                      |
| F-03 · derived from the recorded run data                                                                                                       |                      |

Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

R-01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

cli/codex-gpt-5.5

WHAT WE WOULD LOOK AT FIRST

Review the measured annotation metrics and validator error codes against the declared rubric.

R-02

For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T032642Z**

EXAM FINGERPRINT

**c3b4773c375c10ad989357bb179cffd62151eccd2a58b52254609a54d434be33**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**f041ab53c0abd3fbd8155b14173cb956bb0ccdb8f3abe59c2b0ec260e46b29ed**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**4ad18e29e7607504d657ce0365a3b8ccce22c6c8a54d020104848d9614918076**

sha256 of fr-20260906-ee230324.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**176972c359716e78be2af6b91b61596361f75dc0a95c8cd4d259418d111453ab**

sha256 of fr-20260906-ee230324.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**bc53da7461d040f5d2d0904f4b7f4d2e3c194073c0b2009270a90da6b76314e5**

sha256 of fr-20260906-ee230324.pdf.

RUN EVIDENCE BUNDLE

**ee23032446105d9b04cf546d4eed629c9ef94ea012ec6dadff089ce470be6f6f**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**a0997059eb829c98be90048ef2ed2bbc9debfb0d69af0e51674a13bda168bea1**

The exam's sealed evidence, established before this run.

Measurement history

This is the first recorded run on this exam. There is nothing to compare it against yet, and a single measurement is not a trend.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 9e118309 | cli/codex-gpt-5.5 | 0    | submitted_failed | yes       | Graded        | fail         | 1     | n/a    | 123.6s  |

Failure taxonomy

| CLASS            | COUNT | WHAT IT MEANS                                             |
|------------------|-------|-----------------------------------------------------------|
| submitted_failed | 1     | Submitted annotations rejected by the independent grader. |



Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

ee23032446105d9b04cf546d4eed629c9ef94ea012ec6dadff089ce470be6f6f

RUN ID

live-20260906T032642Z

STARTED

2026-09-06T03:26:42.461624+00:00

FINISHED

2026-09-06T03:28:48.278924+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 1afdb5a2d08b69212096eba50994a7dc6466ddef1a6919962f7c2492c9f8da58 |
| rollouts/9e118309-ded3-4dc1-a44b-dfa232943eda.json | d4978ce7e6060d254b891523327dc5b2850adc74cebf8d1ac01a38bb0722f    |
| spec                                               | a5ceb66bfcdc2a2a4a2d19f8f76a3dc63b4fe81b7eff9e4cba1c5cddbc9e6be2 |
| summary                                            | e81e63889d9fc3bc061f5b278ceb486578d2c3798bd4f1bf1b73f22e61ccc946 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

9e118309 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_failed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 79    |
| 1    | read_file  | TASK.md              | ok     | 732   |
| 2    | read_file  | annotation-task.json | ok     | 4 329 |
| 3    | read_file  | assets/robot.webm    | ok     | 4 416 |
| 4    | read_file  | assets/robot.webm    | ok     | 822   |
| 5    | read_file  | submission.json      | ok     | 74    |
| 6    | write_file | submission.json      | ok     | 64    |
| 7    | run_tests  | —                    | ok     | 186   |
| 8    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (60 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

29 Attempt record 1 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-0524b448 · record digest 6c7ef1ad19e86c26... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

1

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.5s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-0524b448                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:48:31 | Evaluation run started                               |
| 2026-09-06 03:49:18 | Evaluation run finished                              |
| 2026-09-06 03:49:18 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 46.1s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 3dcca802-5fb2-45ca-a14e-1d1f91ba64d6

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034831Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**6c7ef1ad19e86c265db5ca3ae26545f32a3aa7610be592847156999eaad0a613**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**5b350d8a9ef8c2dea17c766be25b277f1a9902f09c55004ce814976b7dd721e8**

sha256 of fr-20260906-0524b448.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**b9d931194681692f8e173ffb2006a188f93655e1a35a2b5c3a94b8719f4e3969**

sha256 of fr-20260906-0524b448.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**a6623097d00bf6d901d28e1401dfadeee8630fb07d69d245d5e86499ffcc93b8**

sha256 of fr-20260906-0524b448.pdf.

RUN EVIDENCE BUNDLE

**0524b4485db92134fa1c14f30a45861cd310c37e2bdabbbebf9dc9447849b8e**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-53c54cfc`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 3dcca802 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 46.1s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>0524b4485db92134fa1c14f3</div> <div>0a45861cd310c37e2bdabbbe</div> <div>bfb9dc9447849b8e</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T034831Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:48:31.132363</div> <div>+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:49:18.626062</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | d457ee12999fc9b4e21e00eb8dde2e4a1dfc1fe08d2a785a1eca358b5cf57328 |
| rollouts/3dcca802-5fb2-45ca-a14e-1d1f91ba64d6.json | 9aaeaf52daedb230cc70992ec757f26d9bb7f05c55d7b77d5a42c965965a4a03 |
| spec                                               | 3470954e0613468dcebase44a03f9010a711eba7eb0f559f0be020d7ed31cc59 |
| summary                                            | 1fc634f45dba17a2641c5b9ab2f16101d570ba8c2bfd8407cafb9a3364be23ad |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.



Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

3dcca802 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

30 Attempt record 2 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-53c54cfc · record digest 04d2e9b4b28fa343... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 50.3s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-53c54cfc                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:47:40 | Evaluation run started                               |
| 2026-09-06 03:48:31 | Evaluation run finished                              |
| 2026-09-06 03:48:31 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 48.9s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 4a993bea-c10e-435d-aff3-ba50480a2767

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.



Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASSEvidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.  
F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTEEvidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.  
F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034740Z

EXAM FINGERPRINT

5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**04d2e9b4b28fa343d7827d6cab7a2a40a2b7d60e2a7ebbe6a19df7d214bb1de4**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**7a250634c4dcd6d835f808b7a08022238f62f6491397b65da761116a8b93c274**

sha256 of fr-20260906-53c54cfc.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**47b30f91355ac68c768eae3a418416b348e4d209a95331c676d8d324d6485d35**

sha256 of fr-20260906-53c54cfc.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**17642c4b347c0342420ac0551f5296f369f18579ca2fd3cd0bee1cfe4abf568c**

sha256 of fr-20260906-53c54cfc.pdf.

RUN EVIDENCE BUNDLE

**53c54cfc62e938a0156792b0b132cc7128196df389015bb743414b71916eae89**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-f1d50a1b. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 4a993bea | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 48.9s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

53c54cfc62e938a0156792b0b132cc7128196df389015bb743414b71916eae89

RUN ID

live-20260906T034740Z

STARTED

2026-09-06T03:47:40.794251+00:00

FINISHED

2026-09-06T03:48:31.103621+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 8d9462283ed11e1026120cbb154e2f73cae167660682ebec38d6f4b474aeaccd |
| rollouts/4a993bea-c10e-435d-aff3-ba50480a2767.json | 7e4e4dd6ef2b9f202d9c5d06f0d92847bf97601428fb3507b4a830f5925bcdd1 |
| spec                                               | 8dd820b1d487bcf18ab9ba44ff9e918b071e6e9897861ce5a3eae8bf03f942dd |
| summary                                            | a5131f1fb6c6770370656b68622cdbd8cadcc513d2f6784717bf03818f17c1cd |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

4a993bea · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (40 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.



31 Attempt record 3 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-6acfa16c · record digest 79cfd5852be5991c... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

1

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.6s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-6acfa16c                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:45:28 | Evaluation run started                               |
| 2026-09-06 03:46:16 | Evaluation run finished                              |
| 2026-09-06 03:46:16 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 46.2s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission fd27f8f1-b62c-4fdf-ab92-c5c160af09b0

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034528Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**79cfd5852be5991c5c7a8e00c6855b01deda4cdd257fa08bb9e1eeb114640a12**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**950283aaa26f92a816fd9efc4e9975a335034c0d5d41035b896c6abca6a82195**

sha256 of `fr-20260906-6acfa16c.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**8d39fc31d679266b6446b2cac720be9e022a1bb6e7d1d0f9e30b1e30ee79d2d5**

sha256 of `fr-20260906-6acfa16c.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**0edc896330dc69040bedc577e998beb2bb84d1cccdacb8102700b335dc4aaa98**

sha256 of `fr-20260906-6acfa16c.pdf`.

RUN EVIDENCE BUNDLE

**6acfa16cea865e71486da13907b4645679217bbad277b3bb48dccbacac73a715**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-ebf6f550`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| fd27f8f1 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 46.2s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                          |                                                    |                                                                |                                                                 |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>6acfa16cea865e71486da13907b4645679217bbad277b3bb48dccbacac73a715</div> | <div>RUN ID</div> <div>live-20260906T034528Z</div> | <div>STARTED</div> <div>2026-09-06T03:45:28.734409+00:00</div> | <div>FINISHED</div> <div>2026-09-06T03:46:16.360509+00:00</div> |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | c8731498247cb21b96695095c4f6f52960ddb696d75d7d422bed09317a4cffd4 |
| rollouts/fd27f8f1-b62c-4fdf-ab92-c5c160af09b0.json | 7eda52fe50278993375a9eaac22341b0a957864798526f5c91bd3cbc0cdbbe8d |
| spec                                               | 1df257c177fa796d21fb943205bdc681a855cb8f6418aaf8461cc3ced4bfda54 |
| summary                                            | e23c46f06cc1cc28d82aaf4bbb2792971bcaf88e3e4106fba7d1c82f3fda0ddd |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

fd27f8f1 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (40 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

32 Attempt record 4 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-81881665 · record digest 6d4694530740ae02... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.2s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-81881665                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:43:14 | Evaluation run started                               |
| 2026-09-06 03:44:01 | Evaluation run finished                              |
| 2026-09-06 03:44:01 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 45.7s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 92967099-321e-4ab4-bed5-bd93705f8a87

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.



Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASSEvidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.  
F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTEEvidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.  
F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034314Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**6d4694530740ae02b13144a46a80351a3fe98fa31d89d3424737da904225ab37**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**c3d2d87505127e20b502533d33999008e1a7398c70a06057901b01bad0807360**

sha256 of fr-20260906-81881665.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**53e32682e877554d9509c94585dc21a7e97ac6425f7ab744175f95491ff64aa5**

sha256 of fr-20260906-81881665.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**969fdc86b03b439e94bd12fb062cd68b73cb7826aa986e58fac0bcd9f281faa4**

sha256 of fr-20260906-81881665.pdf.

RUN EVIDENCE BUNDLE

**818816654e8d5d8e3238328b6adc7fe3b284d30bc79b42e54faf9d15ed62a93e**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

This is the first recorded run on this exam. There is nothing to compare it against yet, and a single measurement is not a trend.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 92967099 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 45.7s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>818816654e8d5d8e3238328b</div> <div>6adc7fe3b284d30bc79b42e5</div> <div>4faf9d15ed62a93e</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T034314Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:43:14.767608</div> <div>+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:44:01.978953</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | a3d8392b88609eaf1b9256c391b4fe40a736c5534e05a53a3bd5e1c7218b2a15 |
| rollouts/92967099-321e-4ab4-bed5-bd93705f8a87.json | 8320cf10e98c58b8992224af7f3e785a2fbaa7cc2b1f5a7647f51ba64b211861 |
| spec                                               | b9045e97deffe09c735400910f6794b6a2732c07b3060d4a2e68795edf3d5a23 |
| summary                                            | 51c9f548b196d6b9f51235fb92cc3038937287689fab50155692c9d2f2b7c809 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

92967099 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

33 Attempt record 5 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-8a3e7827 · record digest 1307b28564b6fd0b... · harness clean · c  
ontainment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vv  
dex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the  
identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation,  
against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence,  
what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation  
workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is  
below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not  
the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for  
acting on them follow it. Elapsed wall clock for the whole engagement: 46.6s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.



Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-8a3e7827                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:46:16 | Evaluation run started                               |
| 2026-09-06 03:47:02 | Evaluation run finished                              |
| 2026-09-06 03:47:02 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 45.2s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 213c95ee-bc6a-4888-a902-66659f03a9fb

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data



## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034616Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**1307b28564b6fd0bdaabd143ff60207f7ea1c12dbfe5c387f37bb3f691d44b3a**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**7595de3db0a6b2543364b61b85d9beb7098e481d79d5e58554dc79543bcc71a**

sha256 of fr-20260906-8a3e7827.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**4ae240aafba9d4f3a02f5366fd31ab371d8d22da5e6c38f9781ccf65c60bff06**

sha256 of fr-20260906-8a3e7827.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**b6310f45e18d45ab0ce3ddda b471d660ed028dd88ecbc2c52d0b80052da9aaf**

sha256 of fr-20260906-8a3e7827.pdf.

RUN EVIDENCE BUNDLE

**8a3e7827b26fac1d4cd4faef f1cb0b639a85e35179b9b46412ae0fc3e7c2d5cb**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69b a0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-6acfa16c. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 213c95ee | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 45.2s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>8a3e7827b26fac1d4cd4faef</div> <div>f1cb0b639a85e35179b9b464</div> <div>12ae0fc3e7c2d5cb</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T034616Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:46:16.391394</div> <div>+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:47:02.978108</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | a7e4095d38a9f08052cde8e234112c50fead24b5884fba472fb960c9377834b4 |
| rollouts/213c95ee-bc6a-4888-a902-66659f03a9fb.json | 02b6e37f34b6cfbf050222f5d1216a7c6e159649a4c1d697b2cc4230ec1eae4  |
| spec                                               | d51be12b12b6d0612fecb18bbc1eb46ac70e1095d391714402df9e52637ec86d |
| summary                                            | 7f0d076c5dd423bf557a61713793cc84659c7ff0a6cf05c15094052fd668276a |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

213c95ee · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

34 Attempt record 6 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-8e07c82f · record digest b6767314873b3ff1... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 42.1s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.



Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-8e07c82f                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:49:18 | Evaluation run started                               |
| 2026-09-06 03:50:00 | Evaluation run finished                              |
| 2026-09-06 03:50:00 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 40.6s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission a02bae7b-c1fa-4f46-a92c-4c104669b287

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.



Verification

RUN ID

live-20260906T034918Z

EXAM FINGERPRINT

5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**b6767314873b3ff13f6ce7f53682b9c383c88196633b2cdc**  
**eae9e044e06b1358**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**9bf658025a304df067cdf5c2bf04d2ed4606a373902406651f3bcec25b9d16a3**

sha256 of fr-20260906-8e07c82f.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**8c73e3b15e0e8dd0d22b3a4047e334cb1403ae7f04c103df639a3a47b3475664**

sha256 of fr-20260906-8e07c82f.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**56031260eea3e54fecaa9c44dfe7e6190aa1efb23e4c37deda9d5362fa4c3d2b**

sha256 of fr-20260906-8e07c82f.pdf.

RUN EVIDENCE BUNDLE

**8e07c82f663d3d954ace1bb6ef1451cf97b5fe2ca137234f61831286a314d256**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-0524b448. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| a02bae7b | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 40.6s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

8e07c82f663d3d954ace1bb6ef1451cf97b5fe2ca137234f61831286a314d256

RUN ID

live-20260906T034918Z

STARTED

2026-09-06T03:49:18.672894+00:00

FINISHED

2026-09-06T03:50:00.796058+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 0d9627d0eda9832e27adf08613b9d08b442aeadc0122cec963ed3e959df91711 |
| rollouts/a02bae7b-c1fa-4f46-a92c-4c104669b287.json | 4870777be623c4a5356354eca6b4e6db5a2bad5aa1c0846aaf9f24985bde3243 |
| spec                                               | 74b8e9c5b2d0c2da0aae8a5f1abbf5d3a95aa9b781c09ff11da4065d9fc8c1ad |
| summary                                            | 789a58840984fa19f706c4204ba0d9fe19f14e7d60b1de20a94dd4dc65636de3 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

a02bae7b · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

35 Attempt record 7 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-a9e8fca4 · record digest c3be15a7ee24dc46... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

1

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 43.1s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-a9e8fca4                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:44:02 | Evaluation run started                               |
| 2026-09-06 03:44:45 | Evaluation run finished                              |
| 2026-09-06 03:44:45 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.



Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 42.1s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 2bf3ef7e-e486-460a-b592-28a5c7a66272

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASSEvidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.  
F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTEEvidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.  
F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034402Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**c3be15a7ee24dc46e4f61be830b3111cbf48792324238e28295020c33354e496**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**dc8081798cf89d53689481317b82df3f3aff82affdbc55fd3102ab56267e3763**

sha256 of fr-20260906-a9e8fca4.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**63b3a02e80674f1a2fd99bf36c8de15e89351f73dc1aa7f38e3c0be22a3fcf7a**

sha256 of fr-20260906-a9e8fca4.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**9dffdaaa1b819b42d853608aece955e8542271a591d16353a7b57616af307cc4**

sha256 of fr-20260906-a9e8fca4.pdf.

RUN EVIDENCE BUNDLE

**a9e8fca4e9276150c0accbf7662615629ba4ec9a2db8cb3d5d2bc8d3b942c166**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-81881665. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.



Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 2bf3ef7e | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 42.1s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                          |                                                    |                                                                |                                                                 |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>a9e8fca4e9276150c0accbf7662615629ba4ec9a2db8cb3d5d2bc8d3b942c166</div> | <div>RUN ID</div> <div>live-20260906T034402Z</div> | <div>STARTED</div> <div>2026-09-06T03:44:02.018538+00:00</div> | <div>FINISHED</div> <div>2026-09-06T03:44:45.138099+00:00</div> |
|----------------------------------------------------------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------|-----------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 8086188369494444346b647d155b435c669d0361749f0546b53a95b4e9c021e1 |
| rollouts/2bf3ef7e-e486-460a-b592-28a5c7a66272.json | 5a794e7b92916444f82a5823eb72c3741416e53df4b30e7885f05168edad1728 |
| spec                                               | 4fdab44ed88a92093dc77a9fea9df2aca9405a83e6e648dde038451b059665   |
| summary                                            | cc40f88643e5341f5b5b34ed3eaa0c0c634da821cb23fe46d144c7e6d3f4f76e |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

2bf3ef7e · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

36 Attempt record 8 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-b9a3013f · record digest 3753722b773d4e4a... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 53.7s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-b9a3013f                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:50:00 | Evaluation run started                               |
| 2026-09-06 03:50:54 | Evaluation run finished                              |
| 2026-09-06 03:50:54 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 52.3s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission db24ff75-e665-4604-88a2-8ded8d8040fd

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T035000Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**3753722b773d4e4ae317ea729b2f76946a55af57cdf71aa5f64727a2c9c1d**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**7d5aa7ddd42b5a604ee0694d2b61fd14f826a9131432326870125e88e4f0d10f**

sha256 of fr-20260906-b9a3013f.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**3295c7f0a213d2472e1680e6e3b1bac09f9dfbaccd4a150654150cd63213ea9c**

sha256 of fr-20260906-b9a3013f.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**276ce7059a195e716851e0a3103c2f246a175cc6089845bc8829724fafcdfda4**

sha256 of fr-20260906-b9a3013f.pdf.

RUN EVIDENCE BUNDLE

**b9a3013f047bfc42b72371c99a4ec83fec7c171d17463ba202382339b065af5f**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-8e07c82f. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| db24ff75 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 52.3s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |



Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>b9a3013f047bfc42b72371c99a4ec83fec7c171d17463ba202382339b065af5f</div></div> | <div><div>RUN ID</div><div>live-20260906T035000Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:50:00.820297+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:50:54.529409+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 7e5075e4534b87c57d6bd7baf13be67c858e3b84b0260b762dafc1b0a265dbac |
| rollouts/db24ff75-e665-4604-88a2-8ded8d8040fd.json | d61c132ca543ea27ea610ce49ce738b0793a31f0af97e91a5bf8ae3bed6b5ef6 |
| spec                                               | 74344ccb07656b19170edb9d4bdb6194d51586a160bea77290800557facac3d1 |
| summary                                            | 2b1c521bc8881099717c301e9b882cfe5eae3d784605cf19f6ced5044a21ba61 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

db24ff75 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

37 Attempt record 9 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-ebf6f550 · record digest 98cf7408f3accbb7... · harness clean · containment clean

Executive summary

MODELS TESTED

1

CERTIFIED PASSES

1

of 1 graded rollouts

SUBMITTED

1 of 1

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 43.5s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-ebf6f550                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:44:45 | Evaluation run started                               |
| 2026-09-06 03:45:28 | Evaluation run finished                              |
| 2026-09-06 03:45:28 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 41.8s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 32dddec6-4012-4a4c-b6e4-09e88b5d2eb3

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034445Z**

EXAM FINGERPRINT

**5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**98cf7408f3accbb7029809751c4723e435e720a77c22a94beb98bf611d5406de**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**dfc00f3e5fa74e4a1c894a43862806d17757207593fd9ab599b9a34411a1ff49**

sha256 of `fr-20260906-ebf6f550.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

**6b42530b2113aa44d34ce20cc8edc4cb73ca5f2985178bbe1e1bc68c0b2069a6**

sha256 of `fr-20260906-ebf6f550.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**0d449376dc9af037e3480f189f0ff234deef3165d5574becb5463d4a28368b07**

sha256 of `fr-20260906-ebf6f550.pdf`.

RUN EVIDENCE BUNDLE

**ebf6f550bf74b06a3a4678956664211e7abea50b29cf1e506f00da694b68b0c0**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-a9e8fca4`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 32dddec6 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 41.8s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

ebf6f550bf74b06a3a4678956664211e7abea50b29cf1e506f00da694b68b0c0

RUN ID

live-20260906T034445Z

STARTED

2026-09-06T03:44:45.196778+00:00

FINISHED

2026-09-06T03:45:28.694294+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | cf418b7f9e9f9eb17bbeab6971c213aa13a7a75a1a10f169dbfdd6e7216a33a1 |
| rollouts/32dddec6-4012-4a4c-b6e4-09e88b5d2eb3.json | 2e7772ae38bd89bee32e5a0b3229a952cd97423e657db336825f56d3c1956209 |
| spec                                               | abfc0caad6f722b6747a29cb847f6fc7994fe417eeeac017240a323f991cbb2b |
| summary                                            | 2ad5c9611c619d38310be652b9ff6345be78bae3f27739c44dbbceb26fea5a0e |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.



Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

32dddec6 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | read_file  | —                    | failed | 35    |
| 1    | list_files | .                    | ok     | 81    |
| 2    | read_file  | TASK.md              | ok     | 520   |
| 3    | read_file  | annotation-task.json | ok     | 3 290 |
| 4    | read_file  | assets/records.json  | ok     | 295   |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

38 Attempt record 10 of 10 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260906-f1d50a1b · record digest 5fec386d9f78f894... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 37.7s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.structured-data-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                    |         |            |                 |
|------------------------------------|---------|------------|-----------------|
| EXAM                               | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.structured-data-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-f1d50a1b                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.structured-data-1 · v0.1.0                           |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:47:03 | Evaluation run started                               |
| 2026-09-06 03:47:40 | Evaluation run finished                              |
| 2026-09-06 03:47:40 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 36.5s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: structured. Rubric: 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 1099a3c1-2c96-43a8-b4f1-494cc773529c

| Score | Items | Valid / invalid | Types     |
|-------|-------|-----------------|-----------|
| 1     | 4     | 4 / 0           | pair, row |

Measured metrics

|                 |   |              |   |               |   |
|-----------------|---|--------------|---|---------------|---|
| annotationCount | 4 | matchedCount | 4 | mismatchCount | 0 |
| pairAccuracy    | 1 | pairF1       | 1 | pairPrecision | 1 |
| pairRecall      | 1 | rowAccuracy  | 1 | rowF1         | 1 |
| rowPrecision    | 1 | rowRecall    | 1 |               |   |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.



Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS          | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|---------------|----------|------------------|----------|--------|--------|------------------|
| pairPrecision | recorded | N/A              | 1        | –      | –      |                  |
| pairRecall    | recorded | N/A              | 1        | –      | –      |                  |
| pairF1        | recorded | N/A              | 1        | –      | –      |                  |
| pairAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| rowPrecision  | recorded | N/A              | 1        | –      | –      |                  |
| rowRecall     | recorded | N/A              | 1        | –      | –      |                  |
| rowF1         | recorded | N/A              | 1        | –      | –      |                  |
| rowAccuracy   | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034703Z

EXAM FINGERPRINT

5d6b3708dba2cdf08218a44997058413d0a3aae738299bce8c346bb2b309fc28

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**5fec386d9f78f894df537589ca1c3a9470ec13dfbf92b360d28240b91023ee65**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**711d32bf0f9e22e2e20947c1af33ab14d29ae090b489602025548a5189a5905c**

sha256 of fr-20260906-f1d50a1b.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**7c232231cd5ebd09f5aa00095c92a041542eaa4596974765ae89661a33061b46**

sha256 of fr-20260906-f1d50a1b.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**73fdfb3aa87fad8bf8c6d46d1b321747c22c8ccf40af2a00dc0a754f02f9c7a6**

sha256 of fr-20260906-f1d50a1b.pdf.

RUN EVIDENCE BUNDLE

**f1d50a1b063c8ea558d069b0ad205c4b0114dce54424fddb5bb70e8136afdb9c**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-8a3e7827. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 1099a3c1 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 36.5s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

f1d50a1b063c8ea558d069b0ad205c4b0114dce54424fddb5bb70e8136afdb9c

RUN ID

live-20260906T034703Z

STARTED

2026-09-06T03:47:03.016667+00:00

FINISHED

2026-09-06T03:47:40.765753+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 0a6e332109508f1455f4be41f5405580cb2f2bb7afa1cc72efd737b53004ab5b |
| rollouts/1099a3c1-2c96-43a8-b4f1-494cc773529c.json | 62447078fc46e9c720dc106c5f05ef1d2ba32d38c181dc2c27b4eef355541edc |
| spec                                               | 1100d460ba39ff97027806033a62e221e9d17962fba6c2cad8d7968682e14b4c |
| summary                                            | 5e8aa04876512290d0374b7a007791275492ed3ab2390b6fcb9cc88bda4f0aa6 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

1099a3c1 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 81    |
| 1    | read_file  | TASK.md              | ok     | 520   |
| 2    | read_file  | annotation-task.json | ok     | 3 290 |
| 3    | read_file  | assets/records.json  | ok     | 295   |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.



39 Attempt record 1 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-09f49314 · record digest ad912cc0b1e87920... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 50.9s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-09f49314                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:36:38 | Evaluation run started                               |
| 2026-09-06 03:37:29 | Evaluation run finished                              |
| 2026-09-06 03:37:29 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 49.7s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission 833cc002-8b9c-4a61-b380-9dfb64ffb4d1

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033638Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**ad912cc0b1e879202be847d6ff47326a8b769d3913c33c89df9263326287cb48**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**ad5b6e0c993e25a02b2373af157a508ef2f1db7571315c56c61412be7ce027d5**

sha256 of fr-20260906-09f49314.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**089d9253a62b714739d03eeec3ca99dd009a401e841a8c005f57ebcbe46be4d8a**

sha256 of fr-20260906-09f49314.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**5d170fffa8a4e9e55d7867737d3292bef0572a0984abc5093ac8165da2b7d63f**

sha256 of fr-20260906-09f49314.pdf.

RUN EVIDENCE BUNDLE

**09f49314723caa06f668b107d6fac6891745e832918d71f15decd6dab8ef3a63**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-1ac877d5. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 833cc002 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 49.7s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                  |                                                                |                                                                                        |                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| <div>RUN BUNDLE DIGEST</div> <div>09f49314723caa06f668b107</div> <div>d6fac6891745e832918d71f1</div> <div>5decd6dab8ef3a63</div> | <div>RUN ID</div> <div>live-</div> <div>20260906T033638Z</div> | <div>STARTED</div> <div>2026-09-</div> <div>06T03:36:38.35095</div> <div>3+00:00</div> | <div>FINISHED</div> <div>2026-09-</div> <div>06T03:37:29.268272</div> <div>+00:00</div> |
|----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | dfb473d6847fddad7b3ab3cd5452eccb312772cbe1cd9084bce9a033a0d608ce |
| rollouts/833cc002-8b9c-4a61-b380-9dfb64ffb4d1.json | 2c202330c0784d7dbe53402c8c06723d1a0088fa26b8d6ea4e4d0cf072bc82a8 |
| spec                                               | b23b22057959fa9e5fe79938a8deea0d44811f359e85796fa77ad74c619fca88 |
| summary                                            | d10c86d61cb638317d5c0c22bfc2c17bc19f1780707f2cc06cb8ad0ea376f7c0 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

833cc002 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

40 Attempt record 2 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-1ac877d5 · record digest b5b01baf361a3647... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 47.8s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-1ac877d5                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:35:50 | Evaluation run started                               |
| 2026-09-06 03:36:38 | Evaluation run finished                              |
| 2026-09-06 03:36:38 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 46.6s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission e77671d0-72a8-4be1-9af2-e17094bc97f4

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.



Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T033550Z**

EXAM FINGERPRINT

**e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**b5b01baf361a364761324340a0e7f776d1de9aa5edea6652a2f9bedfe9f7d379**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**60684e3bebf3ad9a76e7ab4efb206bdbfea6a31946c497d2e0dd64985104a1d5**

sha256 of fr-20260906-1ac877d5.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**b9da290ca6a3604f0eee29d154a16e074b4403f8c1c1699015d076ddee8422a0**

sha256 of fr-20260906-1ac877d5.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**f5013f1eb744fa12b975bc7c0ce192a15f99db7eaba2ac458021815c42d34c28**

sha256 of fr-20260906-1ac877d5.pdf.

RUN EVIDENCE BUNDLE

**1ac877d534c00963c8ada4f370cac9e5c75154c5acce099012af9bc45d43fa02**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-50564102. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| e77671d0 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 46.6s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                                          |                                                                         |                                                                                                |                                                                                                 |
|------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>1ac877d534c00963c8ada4f3</div><div>70cac9e5c75154c5acce0990</div><div>12af9bc45d43fa02</div></div> | <div><div>RUN ID</div><div>live-</div><div>20260906T033550Z</div></div> | <div><div>STARTED</div><div>2026-09-</div><div>06T03:35:50.513425</div><div>+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-</div><div>06T03:36:38.323818</div><div>+00:00</div></div> |
|------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 9707455a69052b97a9ad0801503ac3cfd537830a75ddcc0e5b658e62a74c9ee2 |
| rollouts/e77671d0-72a8-4be1-9af2-e17094bc97f4.json | 9b06114443ea2bb56fe326f5e4b5bdb1fe52323585c6de46ab00f61e833d84d3 |
| spec                                               | b0965ff53c3022f4300bd87a7bafc972fd368636bef53d93e5d592228d30adc5 |
| summary                                            | 84fb78b46ffa4b0b7527addc8b70493acc196c4f615c97b2ed321f957b3a2d1c |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

e77671d0 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (30 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

41 Attempt record 3 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-3876bd90 · record digest 87e2b2d954245b76... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 53.3s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.



Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-3876bd90                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:38:17 | Evaluation run started                               |
| 2026-09-06 03:39:10 | Evaluation run finished                              |
| 2026-09-06 03:39:10 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                  |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                        |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file, write_file, edit_file, run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 52.2s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission 1d73df6b-3d51-4346-a181-2260de6a093d

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data



## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T033817Z**

EXAM FINGERPRINT

**e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**87e2b2d954245b76650d144b5e2ed2cc666cb17d73af83e76c5b3f6a23d0bdbb**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**111248b4785ede0ab4a6230a3fbe05bca486cab8cedb5f52c36625bbc6570539**

sha256 of fr-20260906-3876bd90.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**0a9b492b48fb17e26e4ba589ceb7990f01488214910b319422317732b1a1bce9**

sha256 of fr-20260906-3876bd90.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**8a883d834589ef9b3ea37910df9b5a9db7fb476f2f6aaeef63287405b7299c9d**

sha256 of fr-20260906-3876bd90.pdf.

RUN EVIDENCE BUNDLE

**3876bd9094f4e4b04fe8eb9e3f070591f1162529ade478770fb5df6c6933d7f2**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-5c3a23d2`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 1d73df6b | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 52.2s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

3876bd9094f4e4b04fe8eb9e3f070591f1162529ade478770fb5df6c6933d7f2

RUN ID

live-20260906T033817Z

STARTED

2026-09-06T03:38:17.521191+00:00

FINISHED

2026-09-06T03:39:10.846763+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 99971fac8fce0c4797ae32e54033083865f52fe6c448f4753eed18092a1a791b |
| rollouts/1d73df6b-3d51-4346-a181-2260de6a093d.json | 05aaa73f6e101279f737f376e2ad4346637dba22522cebda79ffcdbe89156509 |
| spec                                               | 2157f2817d9209a32ae87b021686c86d1516b555086eac933b94e6b0a0055f19 |
| summary                                            | f5a9fb761ec05bc7b720197f3aad02471cd1b0e9e6f8b7e9e0235fcf9dbc5be7 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

1d73df6b · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

42 Attempt record 4 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-38dcc889 · record digest 003299dc0eca735f... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 56.8s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.



Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-38dcc889                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:39:59 | Evaluation run started                               |
| 2026-09-06 03:40:55 | Evaluation run finished                              |
| 2026-09-06 03:40:55 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 55.4s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission 4b3a698f-9475-4244-9647-06f3f00849de

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.



Verification

RUN ID

**live-20260906T033959Z**

EXAM FINGERPRINT

**e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**003299dc0eca735f784408e749762c887a90832c70bc66439503b7da10333eee**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**a237b1448bf6ca5148a66e5fc ca425111e6f0ccc0e68ee5c135c309b834e89bc**

sha256 of fr-20260906-38dcc889.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**33fb52c7cee1f525b2ce4543e1b5874a49fefac3972b6a3903fe5a319ef2d306**

sha256 of fr-20260906-38dcc889.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**bd2ba2efffc01b250e5cea9e5dccebacc669eba9d372e726aeef78145ca39b470**

sha256 of fr-20260906-38dcc889.pdf.

RUN EVIDENCE BUNDLE

**38dcc889476e3d91c2b8edbf75fde788dbc9058334eae914b5d71d342b25099**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-8a0a838e`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 4b3a698f | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 55.4s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

38dcc889476e3d91c2b8edbf

75fde788dbc9058334eaec91

4b5d71d342b25099

RUN ID

live-

20260906T033959Z

STARTED

2026-09-

06T03:39:59.179654

+00:00

FINISHED

2026-09-

06T03:40:55.979730

+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | a691bd5c35a45eefce9e2e3f04a8e63f1fd558de573db911c4aca8c097657e13 |
| rollouts/4b3a698f-9475-4244-9647-06f3f00849de.json | 995f47ef8ba816bc6f245eec26c9fd068c59781c3646207c6248b73f8cb3ee6e |
| spec                                               | 9c5a81009e800fc925202b9a7962100ba661b312a2b6a22f6361c1edddc988e8 |
| summary                                            | 7ad8bd4175f1faac1ed82153f8eaff4ab4b2372dca91395ec523cb7c4b05772b |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

4b3a698f · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (30 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

43 Attempt record 5 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-49710bdf · record digest 33104a999a042890... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 42.2s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-49710bdf                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:40:56 | Evaluation run started                               |
| 2026-09-06 03:41:38 | Evaluation run finished                              |
| 2026-09-06 03:41:38 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.



Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 41.0s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission c1e330be-f0c0-4ec6-ba47-a95e5be35337

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 8 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034056Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**33104a999a0428902401b7ffeaa929b2de537af21b71863d2de01b5868fe21f9**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**02798a4ff9fd0bed6935981a88a684fb25fa98389c1270fb55e7988ca5eb7def**

sha256 of fr-20260906-49710bdf.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**5ddf66e79add273c9468b6771cb9a88babe58d117131c347d16847c0527a7f8d**

sha256 of fr-20260906-49710bdf.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**a2274cded968151abed7d3dd7c11abbccc6aa8ba7931ce7b3c04887a5e23ed29**

sha256 of fr-20260906-49710bdf.pdf.

RUN EVIDENCE BUNDLE

**49710bdf081703608e704167b15995be6fb2b20f51e7c16fbd69187ac04cff6e**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-38dcc889. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.



Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| c1e330be | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 41.0s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                    |                                                              |                                                                          |                                                                           |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>49710bdf081703608e704167b15995be6fb2b20f51e7c16fbd69187ac04cff6e</div></div> | <div><div>RUN ID</div><div>live-20260906T034056Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:40:56.007984+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:41:38.220490+00:00</div></div> |
|--------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 587acf8c13e8ea3358f80f4b4645b82f469783e6ec5ffce3c390c23dab29b77b |
| rollouts/c1e330be-f0c0-4ec6-ba47-a95e5be35337.json | 5a58ba4d40a85ffbfcdb374af23ccd5f85c9d2baa3923517659b40e1cdf775be |
| spec                                               | 8f488a702bb5dad5ac6a08c3ad1edc9db628834d13374133d7a4858e84db54da |
| summary                                            | 9a1423646158d85cb07330422825423e5836ff4222ad54e97c19e42276116ff3 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

c1e330be · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | read_file  | submission.json      | ok     | 74    |
| 5    | write_file | submission.json      | ok     | 64    |
| 6    | run_tests  | —                    | ok     | 186   |
| 7    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

44 Attempt record 6 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-50564102 · record digest 827e57535e98985d... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 69.3s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-50564102                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:34:41 | Evaluation run started                               |
| 2026-09-06 03:35:50 | Evaluation run finished                              |
| 2026-09-06 03:35:50 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 67.8s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.



Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission aba75eb9-d017-4674-9ca1-a120727cdcbe

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033441Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**827e57535e98985d0fbe512136942a78163cd6d2123acab263ca597bdd883077**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**d16e46fb207db0de97acde822c8d32c1a9b967591dc5a1ace7b311ba8bff76b3**

sha256 of fr-20260906-50564102.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**810f073afb35dcb525c52cb8cc4de0dc6884b9564f80565382bdfee9e231fd41**

sha256 of fr-20260906-50564102.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**e325d1c3140d40ae78e0e2984aa92b6d6bf37b7bf4e1fe667f521117c3b8ef09**

sha256 of fr-20260906-50564102.pdf.

RUN EVIDENCE BUNDLE

**505641029d69128da173adef3128e24bd83d4bdf3aa69037c91425347c741077**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

This is the first recorded run on this exam. There is nothing to compare it against yet, and a single measurement is not a trend.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| aba75eb9 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 67.8s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |



Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

505641029d69128da173adef3128e24bd83d4bdf3aa69037c91425347c741077

RUN ID

live-20260906T033441Z

STARTED

2026-09-06T03:34:41.159703+00:00

FINISHED

2026-09-06T03:35:50.480786+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 896ab507d322c6f1477128e9477649bf35693452bcc74a1578cc0c0bd223a21b |
| rollouts/aba75eb9-d017-4674-9ca1-a120727cdcbe.json | c874757d2840e697ef85c768a23d821ee109af46b8052d55c85a15a2e3940caa |
| spec                                               | c27b1e900e4b70ddfcf3b4dac422fd8d8f067c26596204e4db23a8a02c828e9a |
| summary                                            | 510ce01cfb0e5b5e247529437f5f817a37933ec4c309b14d8ea7845997cc3b2b |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

aba75eb9 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

45 Attempt record 7 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-5c3a23d2 · record digest ec06840685b25bea... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 48.2s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-5c3a23d2                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:37:29 | Evaluation run started                               |
| 2026-09-06 03:38:17 | Evaluation run finished                              |
| 2026-09-06 03:38:17 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 47.0s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.



Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission b75612cd-7374-46fc-b2bc-d2c290096f6c

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASSEvidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.  
F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTEEvidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.  
F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033729Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**ec06840685b25beaf6b6df8775cc34a377131af1c61026b278359b8f422db04f**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**6f1c9fbb876b4f7dfea80e007e7d82e94dc9dfe625434bfa1ee588565848a7b3**

sha256 of fr-20260906-5c3a23d2.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**2904b5fe87bec5f2fbbd9e22aadb4e3e9905e3f9dc76a53a0190ecd746328513**

sha256 of fr-20260906-5c3a23d2.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**a1f16411b6989cb70f1f96e292802a656a7816a41acac1c1b8f9980d2d1ae54c**

sha256 of fr-20260906-5c3a23d2.pdf.

RUN EVIDENCE BUNDLE

**5c3a23d21b8d3fc5cf93f04a3c3c10f67a2949fa21063adcde12901bcf4670c0**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-09f49314`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| b75612cd | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 47.0s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

5c3a23d21b8d3fc5cf93f04a3c3c10f67a2949fa21063adcde12901bcf4670c0

RUN ID

live-20260906T033729Z

STARTED

2026-09-06T03:37:29.316885+00:00

FINISHED

2026-09-06T03:38:17.492331+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | 056d6e32e105519acdfddcc5100ba8d6125d80167e99bf56093131c4354a1f0f |
| rollouts/b75612cd-7374-46fc-b2bc-d2c290096f6c.json | db2775c08365db4b4f8b66116540cdf3e784b294051a94a475664bd9cce2957a |
| spec                                               | 91d63d1955f3e07faa026ec71ab5b48f1158a02406d4c9d8836d048dcf7d9e6a |
| summary                                            | d7f0f91bf918f9324938033ea0f90f1211c4c8d8425d0db811769ed36c0c7a81 |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.



Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

b75612cd · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (30 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

46 Attempt record 8 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-64a9e783 · record digest 314ad299e8c74d9f... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 46.2s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-64a9e783                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:42:28 | Evaluation run started                               |
| 2026-09-06 03:43:14 | Evaluation run finished                              |
| 2026-09-06 03:43:14 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                  |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                        |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file, write_file, edit_file, run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 44.5s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission 5794d22b-99d5-4fd4-80b4-f71d200abe17

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.



Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T034228Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**314ad299e8c74d9f76a9615eb5054c16c73489e4a87ef0e3506ae7c7bff8993b**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**0cdf0a46f5c38e51e120865d92dc8fec3633fce7819ca0597afc7daccb07d278**

sha256 of fr-20260906-64a9e783.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**f6de30634403c5d8005a53454b162da03108f8a3011dd4ce6d8b922cb634d17a**

sha256 of fr-20260906-64a9e783.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**db5a40b4b71766ef7152d1cf917f62ea400e87355759f3c2d200f60ecf04a81c**

sha256 of fr-20260906-64a9e783.pdf.

RUN EVIDENCE BUNDLE

**64a9e783e55e11f6049b4cfd75fdf23ec361778967b63853c48e9f032ad80a4c**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-80719317`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 5794d22b | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 44.5s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

64a9e783e55e11f6049b4cfd75fdf23ec361778967b63853c48e9f032ad80a4c

RUN ID

live-20260906T034228Z

STARTED

2026-09-06T03:42:28.508135+00:00

FINISHED

2026-09-06T03:43:14.720925+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | f50800e611b3c2267657257a3d51ea567e8bd405e88b02c6a50780d45c7b6590 |
| rollouts/5794d22b-99d5-4fd4-80b4-f71d200abe17.json | adb613007d69eeb8c39746659e85c77a7bddbc08637cfc709f487635756e3313 |
| spec                                               | 50c420015e6531198a100c8c5a273d6450f207c37f924ef161119f0e3941dc82 |
| summary                                            | c090c859878b094d5db6d6eb9ccae7e37bf5e4f3c174fd2eeb422e0a03a0260e |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

5794d22b · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (30 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.



47 Attempt record 9 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-80719317 · record digest cb2f2d25f460da6d... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 50.2s.

Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-80719317                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:41:38 | Evaluation run started                               |
| 2026-09-06 03:42:28 | Evaluation run finished                              |
| 2026-09-06 03:42:28 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                                       |                                         |                                              |                                                                                                     |
|---------------------------------------|-----------------------------------------|----------------------------------------------|-----------------------------------------------------------------------------------------------------|
| <div>RUNTIME</div> <div>docker</div>  | <div>STEP LIMIT</div> <div>24</div>     | <div>WALL-CLOCK LIMIT</div> <div>3000s</div> | <div>MAX OUTPUT TOKENS</div> <div>2048</div>                                                        |
| <div>TEMPERATURE</div> <div>0.2</div> | <div>RETRY POLICY</div> <div>none</div> | <div>COMMAND BUDGET</div> <div>0</div>       | <div>TOOLS OFFERED</div> <div>list_files, read_file, write_file, edit_file, run_tests, submit</div> |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 48.9s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission 23cb2ab2-cd1b-4de3-8971-270b714f3a0b

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.



Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### **R-01** For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

**live-20260906T034138Z**

EXAM FINGERPRINT

**e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3**

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**cb2f2d25f460da6d9620639467f41d4824c1c8013c8e7cd679da84a34be70cec**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**9c3d492317fc7f8a785d916dab6ed162df5c3c5d5122aa11097ecc56b5e1e6c2**

sha256 of fr-20260906-80719317.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**7141b0e8552d03d50fd2e2345fefac39562c11a1baeb862a72b1c43fc291b52c**

sha256 of fr-20260906-80719317.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**e25df548d6ceb8211fe096cb16104846c893e2c93901f8cad38dedbb96319883**

sha256 of fr-20260906-80719317.pdf.

RUN EVIDENCE BUNDLE

**807193176c24bb326729ffbea7c36b332ecfd09a803788790e51d6fecd0d4278**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: fr-20260906-49710bdf. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| 23cb2ab2 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 48.9s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

|                                                                                                                     |                                                              |                                                                          |                                                                           |
|---------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div><div>RUN BUNDLE DIGEST</div><div>807193176c24bb326729ffbea7c36b332ecfd09a803788790e51d6fecdd0d4278</div></div> | <div><div>RUN ID</div><div>live-20260906T034138Z</div></div> | <div><div>STARTED</div><div>2026-09-06T03:41:38.254768+00:00</div></div> | <div><div>FINISHED</div><div>2026-09-06T03:42:28.474548+00:00</div></div> |
|---------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------|--------------------------------------------------------------------------|---------------------------------------------------------------------------|

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | b71460419411681f8b3a0813e7b614455846a3165750b0fff4c2e827cf809bb5 |
| rollouts/23cb2ab2-cd1b-4de3-8971-270b714f3a0b.json | 10653c8138df1cf653d5eccd47209297ab820e74429055838e67ff1a4f3b694f |
| spec                                               | c7a166da965ac1cc1da17a3394017cfb087b5bb45bb986e0a317a30993243b55 |
| summary                                            | ea4b388f9efd8b6130d64935616e1444c18def89b6258c7cf226ce16c27f0fcc |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

23cb2ab2 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

48 Attempt record 10 of 10 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260906-8a0a838e · record digest 40b107e6f5a2bc46... · harness clean · containment clean

Executive summary

|                                       |                                                                          |                                                                                      |                                                                           |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| <div>MODELS TESTED</div> <div>1</div> | <div>CERTIFIED PASSES</div> <div>1</div> <div>of 1 graded rollouts</div> | <div>SUBMITTED</div> <div>1 of 1</div> <div>graded rollouts that called submit</div> | <div>NOT GRADED — LANE ERRORS</div> <div>0</div> <div>of 1 attempts</div> |
|---------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------|

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.text-labeling-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 1

graded rollouts passed the independent hidden tests and were certified

1 of 1 graded rollouts passed. cli/codex-gpt-5.5 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 1 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages.

Key findings

F-01 cli/codex-gpt-5.5 passed the independent hidden tests

F-02 Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 48.3s.



Exam definition

**In scope.** The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 3000s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

**Out of scope.** General model quality, performance on other task families, prompt-tuning on the lane's behalf, and any comparison this run's sample size cannot support. A lane's API failure is reported as infrastructure, never as model capability.

**The instrument.** Exam `vvdex.annotation.text-labeling-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

|                                  |         |            |                 |
|----------------------------------|---------|------------|-----------------|
| EXAM                             | VERSION | FAMILY     | ROLLOUTS        |
| vvdex.annotation.text-labeling-1 | 0.1.0   | annotation | 1 (1 per model) |

Where the defect came from

This exam has no upstream repository to credit: the defect, the reference solution and the hidden suite are VVDex's own.

Certification evidence

Why this exam is trustworthy. Every pillar below was established before any model saw the exam, loaded from the sealed evidence matching this run's exam fingerprint, and unchanged by this evaluation.

FROZEN BASELINE

FAIL

expected — an empty submission over 2 run(s) must not pass

→

GOLD / REFERENCE

PASS

expected — the reference solution over 2 run(s) must pass

GRADER CONTROLS

19 / 19

The grader accepted the reference and rejected every near-miss.

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

Sealed 2026-09-06 over 13 artifact(s).

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

Full value in Appendix B; the contract digest this run was executed against.

Evaluation configuration

|                |                                                                       |
|----------------|-----------------------------------------------------------------------|
| REPORT         | fr-20260906-8a0a838e                                                  |
| ISSUED         | 2026-09-06                                                            |
| SUBJECT EXAM   | vvdex.annotation.text-labeling-1 · v0.1.0                             |
| FAMILY         | annotation                                                            |
| PREPARED BY    | VVDex Forge engine vvdex-env 0.1.0                                    |
| CLASSIFICATION | Independent model evaluation report — independently verifiable        |
| CERTIFICATION  | Exam certification: recorded and passing (see Certification evidence) |
| STATUS         | FINAL · report sealed against its own digest                          |

Timeline

| WHEN (UTC)          | EVENT                                                |
|---------------------|------------------------------------------------------|
| 2026-09-06 00:01:41 | Exam evidence sealed (before any model saw the exam) |
| 2026-09-06 03:39:10 | Evaluation run started                               |
| 2026-09-06 03:39:59 | Evaluation run finished                              |
| 2026-09-06 03:39:59 | Report generated from the sealed run evidence        |

Models under test

| MODEL                                             | PROVIDER | ENDPOINT      | BILLING                                            |
|---------------------------------------------------|----------|---------------|----------------------------------------------------|
| GPT-5.5 via Codex CLI (subscription, image input) | cli      | local runtime | operator subscription, flat — not billed per token |

A customer endpoint is recorded by origin only — never its port, its path, or its key.

Limits

|                    |                      |                           |                                                                                        |
|--------------------|----------------------|---------------------------|----------------------------------------------------------------------------------------|
| RUNTIME<br>docker  | STEP LIMIT<br>24     | WALL-CLOCK LIMIT<br>3000s | MAX OUTPUT TOKENS<br>2048                                                              |
| TEMPERATURE<br>0.2 | RETRY POLICY<br>none | COMMAND BUDGET<br>0       | TOOLS OFFERED<br>list_files, read_file,<br>write_file, edit_file,<br>run_tests, submit |

The evaluated model never saw the hidden tests or the reference solution, and could not change its result. Grading ran in a separate step, after submission, on a container with no network.

Model outcomes

**Denominators.** Attempts requested: 1. Graded (evaluable) rollouts: 1. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

| MODEL                                                                     | ATTEMPTS | GRADED | PASSED | MODEL FAILS | LANE ERRORS | SUBMITTED | MEAN TIME | TYPICAL DEEPEST STAGE |
|---------------------------------------------------------------------------|----------|--------|--------|-------------|-------------|-----------|-----------|-----------------------|
| GPT-5.5 via Codex CLI<br>(subscription, image input)<br>cli/codex-gpt-5.5 | 1        | 1      | 1 / 1  | 0           | 0           | 1 / 1     | 47.0s     | Passed                |

Every rollout, with its own deepest stage and grade, is in Appendix A.

Success rate, confidence and pass@k

Every rate on this page is a measurement of a finite number of rollouts, so every rate carries the interval that measurement actually supports. n is the evaluable count — attempts minus rollouts the lane ended before grading; those are listed under "not graded" and sit in no rate, interval or pass@k.

| MODEL             | ATTEMPTS | N | NOT GRADED | PASSES | RATE   | 95% INTERVAL | SEEDS | PASS@1 |
|-------------------|----------|---|------------|--------|--------|--------------|-------|--------|
| cli/codex-gpt-5.5 | 1        | 1 | 0          | 1      | 100.0% | [21%, 100%]  | 0     | 100.0% |

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; the sealed record carries every k up to n.

Model comparison

One model ran in this record, so there is nothing to compare it with. A comparison table across models appears here when a run covers more than one.

Stage funnel

Recorded annotation actions and independent grader verdicts. Missing capabilities are N/A; each stage is measured independently.

| STAGE    | CLI:CLI/CODEX-GPT-5.5 |
|----------|-----------------------|
| Inspect  | 1 / 1                 |
| Annotate | 1 / 1                 |
| Validate | 1 / 1                 |
| Submit   | 1 / 1                 |
| Graded   | 1 / 1                 |
| Passed   | 1 / 1                 |

Deepest stage reached

| MODEL                 | DEEPEST (ANY ATTEMPT) | TYPICAL (MOST ATTEMPTS) | ANNOTATION QA PASSES | NOT GRADED |
|-----------------------|-----------------------|-------------------------|----------------------|------------|
| cli:cli/codex-gpt-5.5 | Passed                | Passed                  | 1 / 1                | 0          |

The matrix counts evaluable annotation attempts; lane errors are shown separately.  
Successful recorded actions establish each stage independently.

Annotation evidence

Modality: text. Rubric: 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission a52c7757-b6e1-4da9-a963-f3203a3f9cc5

| Score | Items | Valid / invalid | Types                |
|-------|-------|-----------------|----------------------|
| 1     | 3     | 3 / 0           | classification, span |

Measured metrics

|                         |   |                        |   |                  |   |
|-------------------------|---|------------------------|---|------------------|---|
| annotationCount         | 3 | classificationAccuracy | 1 | classificationF1 | 1 |
| classificationPrecision | 1 | classificationRecall   | 1 | matchedCount     | 3 |
| mismatchCount           | 0 | spanExactF1            | 1 | spanF1           | 1 |
| spanOverlapIoU          | 1 | spanPrecision          | 1 | spanRecall       | 1 |

Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.

Validation error codes

None recorded

Disagreement

|                  |              |
|------------------|--------------|
| Agreement review | Not measured |
|------------------|--------------|

Measured grader aggregates only. Missing metrics or disagreement are not measured, never zero. Reference answers and item-level labels are excluded. Certification and the evidence digest are recorded in this report's verification sections.

Quality axes

Each axis was measured inside the grade box, on the submitted workspace, by the same deterministic script for every rollout. The reward is the conjunction of the required axes; the score is a weighted reading beside it, not the gate.

| AXIS                    | GATE     | ROLLOUTS PASSING | MEASURED | BUDGET | WEIGHT | WHAT IT MEASURES |
|-------------------------|----------|------------------|----------|--------|--------|------------------|
| classificationPrecision | recorded | N/A              | 1        | –      | –      |                  |
| classificationRecall    | recorded | N/A              | 1        | –      | –      |                  |
| classificationF1        | recorded | N/A              | 1        | –      | –      |                  |
| classificationAccuracy  | recorded | N/A              | 1        | –      | –      |                  |
| spanPrecision           | recorded | N/A              | 1        | –      | –      |                  |
| spanRecall              | recorded | N/A              | 1        | –      | –      |                  |
| spanF1                  | recorded | N/A              | 1        | –      | –      |                  |
| spanOverlapIoU          | recorded | N/A              | 1        | –      | –      |                  |
| spanExactF1             | recorded | N/A              | 1        | –      | –      |                  |

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

| MODEL                 | N | STEPS USED | RAN OUT OF STEPS | COMMANDS | CHANGED THE TREE | REFUSED (OVER BUDGET) | REVISITS | DEAD ENDS / ROLLOUT |
|-----------------------|---|------------|------------------|----------|------------------|-----------------------|----------|---------------------|
| cli:cli/codex-gpt-5.5 | 1 | 7 / 24     | 0                | 0 / 0    | 0                | 0                     | 0        | 0                   |

Each column is defined in Appendix D.

Hallucination & overclaim

Three measurements, each with a stated definition. None of them is a judgement of the model's prose by another model — every number below comes from comparing what the model wrote against what was on disk, or against what the independent grader returned.

| MODEL                 | N | GROUNDING FAILURES | KINDS | ROLLOUTS AFFECTED | OVERCLAIMS | OVERCLAIM RATE | RE-READS |
|-----------------------|---|--------------------|-------|-------------------|------------|----------------|----------|
| cli:cli/codex-gpt-5.5 | 1 | 0                  | none  | 0 / 1             | 0 / 1      | 0.0%           | 0        |

No grounding failures and no false pass claims were detected across these 1 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.



Findings

Every finding below is derived from the recorded data of this run — none is an opinion.  
Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA  
infrastructure, not capability · NOTE a property of the measurement.

F-01

PASS

cli/codex-gpt-5.5 passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

1 of 1 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02

NOTE

Statistical floor — 1 rollout(s) is below the 10-rollout rate floor

Severity: NOTE

Evidence: Appendix D

The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.

F-02 · derived from the recorded run data

## Recommendations

One recommendation per observed failure mode, addressed to the lane it affects. Each traces back to a finding above and to the raw trace in Appendix A.

### R-01 For anyone reading the percentages

Commission a run at  $n \geq 10$  rollouts per lane before treating any rate here as a rate; this run's per-rollout outcomes are exact, its percentages are not yet stable.

Verification

RUN ID

live-20260906T033910Z

EXAM FINGERPRINT

e1abe51398b2d950a3415678aaf8ec796420dfb05ea89b3b64afd33c1e345fd3

The contract digest this run was executed against.

Provenance of this chapter

This chapter was rendered by the campaign generator from the sealed evaluation record; it is not the standalone report reproduced byte for byte. Verify the standalone report with its own digest, verify the record with `vvdex-env records verify`, and verify this campaign document as a whole.

RELATIONSHIP

**campaign-rendered chapter**

Derived from the same sealed evaluation record as the standalone report; these bytes are authenticated by the campaign artifact that contains them.

SEALED RECORD DIGEST

**40b107e6f5a2bc46cb85916a0758419edb7567c61980979e348def9a35e3b9b9**

sha256 over the record's canonical JSON — the identity every renderer works from.

SOURCE RECORD FILE

**31486a85b5d77190a8052c545a8600fa9f3e9bd70013283beeec1c99c33d196cf**

sha256 of fr-20260906-8a0a838e.json as written beside the standalone report.

SOURCE STANDALONE REPORT

**9095495e54e487d6ee5e170e216260f0fb9c393af3ca08ea1b607499148275fc**

sha256 of fr-20260906-8a0a838e.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

**3edbff019596d93f3d51325b3fadc742f4d5450cea5ab0e1d39348d38ef189cf**

sha256 of fr-20260906-8a0a838e.pdf.

RUN EVIDENCE BUNDLE

**8a0a838e21b8c88618dc3fa26b6c2468d448b11301d1667d12a677ebfffd5fb6f**

The digest the run's own bundle computed over itself.

CERTIFIED EVIDENCE SEAL

**0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282**

The exam's sealed evidence, established before this run.

Measurement history

Previous run on this exam: `fr-20260906-3876bd90`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

The full artifact-by-artifact digest table, and the reproduction protocol, are in Appendix B.

Appendix A — every rollout

| ROLLOUT  | MODEL             | SEED | OUTCOME          | SUBMITTED | DEEPEST STAGE | HIDDEN TESTS | FILES | TOKENS | ELAPSED |
|----------|-------------------|------|------------------|-----------|---------------|--------------|-------|--------|---------|
| a52c7757 | cli/codex-gpt-5.5 | 0    | submitted_passed | yes       | Passed        | pass         | 1     | n/a    | 47.0s   |

Failure taxonomy

| CLASS                             | COUNT | WHAT IT MEANS |
|-----------------------------------|-------|---------------|
| No failures recorded in this run. |       |               |

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

8a0a838e21b8c88618dc3fa2

6b6c2468d448b11301d1667d

12a677ebffd5fb6f

RUN ID

live-

20260906T033910Z

STARTED

2026-09-

06T03:39:10.868635

+00:00

FINISHED

2026-09-

06T03:39:59.152968

+00:00

Artifact digests

| ARTIFACT                                           | SHA-256                                                          |
|----------------------------------------------------|------------------------------------------------------------------|
| publicSummary                                      | e6ed9a88c44d7735a7df49b2b11c3ffbf359f2fbcf881bc67ff6bb50db8abcc8 |
| rollouts/a52c7757-b6e1-4da9-a963-f3203a3f9cc5.json | 19623f1df5d658d1d65c4e4ef6ae0b4539fbaead2852244c09d453a5fde46984 |
| spec                                               | 274f44e8dc4e18786f6e99191f04df53878bacceb21f4fcd2759202ea4ba50bc |
| summary                                            | 5ec1bfb3f7698a9925495fb0ba551bc9dda3b30826008edc867895bd473829df |

Reproducing this run

Every rollout can be re-derived: the exam is frozen (fingerprint above), the grader ran in the recorded container image, and each rollout's full trace — every tool call, every result, every word — is in the evidence bundle at its rollouts/<id>.json path. A re-run uses the same job shape: the exam id, your endpoint, n, and the runtime. The record id names this run forever.

Appendix C — redacted transcripts

For each rollout: the tool calls it made, in order — which tool, on which of the model's own paths, whether the call succeeded, and how many characters came back. What a tool returned is not here, and neither is the submission: on this exam the submitted content is the graded answer, so no diff, no answer value and no model claim is printed. A request for the reference solution or the hidden tests is recorded as a request; the path it asked for is replaced by a label. See Appendix D, "What this report withholds and why".

Disclosure class applied to this exam: answer-key disclosure (ownership vvdex\_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

a52c7757 · cli:cli/codex-gpt-5.5 · seed 0 · submitted\_passed

| STEP | TOOL       | PATH                 | RESULT | CHARS |
|------|------------|----------------------|--------|-------|
| 0    | list_files | .                    | ok     | 80    |
| 1    | read_file  | TASK.md              | ok     | 409   |
| 2    | read_file  | annotation-task.json | ok     | 2 826 |
| 3    | read_file  | assets/message.txt   | ok     | 69    |
| 4    | write_file | submission.json      | ok     | 64    |
| 5    | run_tests  | —                    | ok     | 186   |
| 6    | submit     | —                    | ok     | 85    |

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

## 49 Evidence

Every number above is recomputable from the sealed records below; each record verifies itself with `vvdex-env records verify`, and this document was regenerated from them with `vvdex-env records campaign`.

- `vvdex.annotation.image-object-1 → fr-20260906-2d00c77f` · recordDigest 648e95b3d571109148c78137d9202e454bec7f03164f20e1ac4d24ca823395e0
- `vvdex.annotation.image-object-1 → fr-20260906-470625ed` · recordDigest d1c1e3f5eaa431b0b2ae53bccae13895768f3d52b8a7d98ddb2f7f1ee6a53e74
- `vvdex.annotation.image-object-1 → fr-20260906-4a88a9bf` · recordDigest 6947e974ea555f5a7d8f4664650e33c2badab9922358dbac8ee209231413d495
- `vvdex.annotation.image-object-1 → fr-20260906-62fb1341` · recordDigest 7a95898a5f2cd0f0b73c4e885cbbff22d2d90754e7602bb646a68df5f8f5a1e6
- `vvdex.annotation.image-object-1 → fr-20260906-6dc38b98` · recordDigest a66b51f5da540d99f9ecf6ff3602d204399934d919da8d7a9859684c63d583c9
- `vvdex.annotation.image-object-1 → fr-20260906-884851f3` · recordDigest 37a729b8483f10b49bdd80e7ba918abe57e5ddc3d6ba8be593d23d816e082d9c
- `vvdex.annotation.image-object-1 → fr-20260906-ad6f18b9` · recordDigest ad3ad73c931070d584cb6d9cb4c645a177b416c9758efe1b62d1bfee32b293b6
- `vvdex.annotation.image-object-1 → fr-20260906-ae6f86705` · recordDigest cde2fc42622e4ef27c1a9505f9a81637fcf0e48bb634045dc0dc9f33e4f87ba5
- `vvdex.annotation.image-object-1 → fr-20260906-af2506a7` · recordDigest b7b6915c0f62d8d5942e7d4d50d86b3967f5e9e4ab77908f010e7f79ecba5497
- `vvdex.annotation.image-object-1 → fr-20260906-fd180ec1` · recordDigest 6a5c91191a02c8ada9bb6fdd380d95e2df6559de255866d09d3af7b2cec1eb23
- `vvdex.annotation.robot-video-1 → fr-20260906-08cf41e3` · recordDigest 4cb0e73aa85083f8e0ef903fd4d28f277a177b2774a66767e5635c84b75f8213
- `vvdex.annotation.robot-video-1 → fr-20260906-12f8c6af` · recordDigest 5fdd4fe7918f8689e1ad975912f7241d661247fdccbc3b30aca8b96499b7a427
- `vvdex.annotation.robot-video-1 → fr-20260906-55c7c9bc` · recordDigest c50dca52eb4e3d7351d38238b8f67be3139846d39eae70d6fb9f3e7edef1d8da
- `vvdex.annotation.robot-video-1 → fr-20260906-57451e50` · recordDigest bde0ebe1b4f14b7989a966b5a11157dc19cb88391c5937324b23d81d9da1a61b
- `vvdex.annotation.robot-video-1 → fr-20260906-5921101f` · recordDigest cafc99313b275331a9ae86c70be3f042b7ab7b8d0b442a9d5100f51c3ba81e93
- `vvdex.annotation.robot-video-1 → fr-20260906-7b7f8fb2` · recordDigest 1879a1fa6919d32cd484a4e39530ec262f54bdc00e7b0b1015ed4d2032245145
- `vvdex.annotation.robot-video-1 → fr-20260906-7f86cd6e` · recordDigest 6637ce529278ae9c381a7407303cca0aa6aadfa60671a0efdedf2ba76e576f18
- `vvdex.annotation.robot-video-1 → fr-20260906-c5386852` · recordDigest 014e24edff39b7c34708f0bc5e8f1c826ba7e8ec940bf705cbb0f96f22d7a832
- `vvdex.annotation.robot-video-1 → fr-20260906-ea40e0a7` · recordDigest 6b7b5285b5fc6dd7c66338dc973344ad8a80f87314711b7aefe95f783f8ad1b5
- `vvdex.annotation.robot-video-1 → fr-20260906-ee230324` · recordDigest f041ab53c0abd3fbd8155b14173cb956bb0ccdb8f3abe59c2b0ec260e46b29ed
- `vvdex.annotation.structured-data-1 → fr-20260906-0524b448` · recordDigest 6c7ef1ad19e86c265db5ca3ae26545f32a3aa7610be592847156999eaa0a613
- `vvdex.annotation.structured-data-1 → fr-20260906-53c54cfc` · recordDigest 04d2e9b4b28fa343d7827d6cab7a2a40a2b7d60e2a7ebbe6a19df7d214bb1de4
- `vvdex.annotation.structured-data-1 → fr-20260906-6acfa16c` · recordDigest 79cfd5852be5991c5c7a8e00c6855b01deda4cdd257fa08bb9e1eeb114640a12
- `vvdex.annotation.structured-data-1 → fr-20260906-81881665` · recordDigest 6d4694530740ae02b13144a46a80351a3fe98fa31d89d3424737da904225ab37
- `vvdex.annotation.structured-data-1 → fr-20260906-8a3e7827` · recordDigest 1307b28564b6fd0bdaabd143ff60207f7ea1c12dbfe5c387f37bb3f691d44b3a
- `vvdex.annotation.structured-data-1 → fr-20260906-8e07c82f` · recordDigest b6767314873b3ff13f6ce7f53682b9c383c88196633b2cdcae9e044e06b1358
- `vvdex.annotation.structured-data-1 → fr-20260906-a9e8fca4` · recordDigest c3be15a7ee24dc46e4f61be830b3111cbf48792324238e28295020c33354e496

- vvdex.annotation.structured-data-1 → fr-20260906-b9a3013f · recordDigest 3753722b773d4e4ae317ea729b2f76946a55af57cdf71aa5f64727a2c9c1d
- vvdex.annotation.structured-data-1 → fr-20260906-ebf6f550 · recordDigest 98cf7408f3accbb7029809751c4723e435e720a77c22a94beb98bf611d5406de
- vvdex.annotation.structured-data-1 → fr-20260906-f1d50a1b · recordDigest 5fec386d9f78f894df537589ca1c3a9470ec13dfbf92b360d28240b91023ee65
- vvdex.annotation.text-labeling-1 → fr-20260906-09f49314 · recordDigest ad912cc0b1e879202be847d6ff47326a8b769d3913c33c89df9263326287cb48
- vvdex.annotation.text-labeling-1 → fr-20260906-1ac877d5 · recordDigest b5b01baf361a364761324340a0e7f776d1de9aa5eda6652a2f9bedfe9f7d379
- vvdex.annotation.text-labeling-1 → fr-20260906-3876bd90 · recordDigest 87e2b2d954245b76650d144b5e2ed2cc666cb17d73af83e76c5b3f6a23d0bddd
- vvdex.annotation.text-labeling-1 → fr-20260906-38dcc889 · recordDigest 003299dc0eca735f784408e749762c887a90832c70bc66439503b7da10333eee
- vvdex.annotation.text-labeling-1 → fr-20260906-49710bdf · recordDigest 33104a999a0428902401b7ffeaa929b2de537af21b71863d2de01b5868fe21f9
- vvdex.annotation.text-labeling-1 → fr-20260906-50564102 · recordDigest 827e57535e98985d0fbe512136942a78163cd6d2123acab263ca597bdd883077
- vvdex.annotation.text-labeling-1 → fr-20260906-5c3a23d2 · recordDigest ec06840685b25beaf6b6ddf8775cc34a377131af1c61026b278359b8f422db04f
- vvdex.annotation.text-labeling-1 → fr-20260906-64a9e783 · recordDigest 314ad299e8c74d9f76a9615eb5054c16c73489e4a87ef0e3506ae7c7bff8993b
- vvdex.annotation.text-labeling-1 → fr-20260906-80719317 · recordDigest cb2f2d25f460da6d9620639467f41d4824c1c8013c8e7cd679da84a34be70cec
- vvdex.annotation.text-labeling-1 → fr-20260906-8a0a838e · recordDigest 40b107e6f5a2bc46cb85916a0758419edb7567c61980979e348def9a35e3b9b9



---

## 50 Attestation

---

Certified results appear only where evidence proves them. Withheld, invalid and lane-error cells are never presented as model results, and no historical evidence was modified in producing this document. Counts are exact; below ten rollouts per cell no rate or ranking beyond the counts is asserted.

© VVDex. All rights reserved. Public materials are provided for viewing, evaluation and verification. Reuse, redistribution, derivative benchmark creation, training use, republication or commercial use requires written permission unless a specific file states otherwise. Full terms: [vvdexops.com/terms/](https://vvdexops.com/terms/). Third-party open source reproduced in an exam keeps its own licence.