

Codex and Claude Sonnet CLIs — text and structured, ten attempts each

2 fixed tasks · 2 sealed attempt records · 40
graded attempts · 40 certified passes

CURRENT
CAMPAIGN REPORT
2026-09-07

Evaluation scope:
cli/codex
cli/claude-sonnet

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

fc-6d6679dea47a · 2 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	4
03.	Environment families	5
04.	Campaign aggregate across these two exams	6
05.	Harness status	7
06.	Cells excluded from capability	8
07.	What the outcomes mean	9
08.	What this report withholds, and why	10
09.	Attempt record 1 of 1 · structured-data-1	11
10.	Attempt record 1 of 1 · text-labeling-1	41
11.	Evidence	71
12.	Attestation	72

01 Executive summary

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

FIXED TASKS 2	SEALED RECORDS 2	MODEL LANES 2	ATTEMPTS 40
GRADED ATTEMPTS 40	CERTIFIED PASSES 40	LANE ERRORS (NOT GRADED) 0	WITHHELD / INVALID 0

Purpose

This report presents a controlled evaluation of 2 model lane(s) against 2 fixed certified annotation tasks, one sealed attempt record 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 two fixed annotation tasks with one sealed attempt record on each task (2 sealed attempts in total). The sealed tasks cover structured, text inputs. 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

40

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-6d6679dea47a · generated from 2 verified sealed record(s)

Key findings across the campaign

- structured-data-1 · cli/codex · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · cli/claude-sonnet · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/codex · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/claude-sonnet · **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 · STRUCTURED-DATA-1 VVDEX.ANNOTATION.STRUCTURED-DATA-1	ANNOTATION · TEXT-LABELING-1 VVDEX.ANNOTATION.TEXT-LABELING-1
cli/claude-sonnet	10/10	10/10
cli/codex	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 two 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	vvdex.annotation.structured-data-1	Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.
Multimodal annotation	vvdex.annotation.text-labeling-1	An invented message is reviewed for classification and entity spans with consistent offsets.

04 Campaign aggregate across these two exams

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
cli/claude-sonnet	20	20	0	0	0
cli/codex	20	20	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 two 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
structured-data-1	answer_key	Applied consistently to 1 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 1 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 1 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260907-cd08adce · record digest 3596e3de9ba226bc... · harness clean · containment clean

Executive summary

MODELS TESTED

2

CERTIFIED PASSES

20

of 20 graded rollouts

SUBMITTED

20 of 20

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 20 attempts

Purpose

This report presents a controlled evaluation of 2 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

20 of 20

graded rollouts passed the independent hidden tests and were certified

20 of 20 graded rollouts passed. cli/claude-sonnet, cli/codex completed the applicable annotation workflow and were accepted by the independent annotation grader.

Key findings

F-01 cli/claude-sonnet, cli/codex passed the independent hidden tests

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: 787.7s.

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 14520s of wall clock, at 10 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	20 (10 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-20260907-cd08adce
ISSUED	2026-09-07
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)
2026-09-07 09:02:22	Evaluation run started
2026-09-07 09:15:29	Evaluation run finished
2026-09-07 09:15:29	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Claude Code CLI (Sonnet, subscription)	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 docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED 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: 20. Graded (evaluable) rollouts: 20. 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
Codex CLI (GPT, subscription) cli/codex	10	10	10 / 10	0	0	10 / 10	56.9s	Passed
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	10	10	10 / 10	0	0	10 / 10	20.2s	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	PASS@3	PASS@5	PASS@10
cli/codex	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
cli/claude-sonnet	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1, 3, 5, 10; the sealed record carries every k up to n.

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/codex	100.0%	[72%, 100%]	n/a	n/a	56.9s	0.0%	none
cli/claude-sonnet	100.0%	[72%, 100%]	n/a	n/a	20.2s	0.0%	none

Tokens and cost read n/a where a lane reports no telemetry (a flat-rate CLI) — an absent measurement, not zero. pass@1 and its interval are over evaluable rollouts only.

This is not a leaderboard. Not separated (intervals overlap): cli/codex vs cli/claude-sonnet. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	CLI:CLI/CODEX	CLI:CLI/CLAUDE-SONNET
Inspect	10 / 10	10 / 10
Annotate	10 / 10	10 / 10
Validate	10 / 10	10 / 10
Submit	10 / 10	10 / 10
Graded	10 / 10	10 / 10
Passed	10 / 10	10 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/codex	Passed	Passed	10 / 10	0
cli:cli/claude-sonnet	Passed	Passed	10 / 10	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 b8eda0e7-467a-47b0-a562-e998a3c58062

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
------------------	--------------

Submission 1477727f-dbc3-450e-b2ed-e82f6d69d655

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
------------------	--------------

Submission e59a52e2-a69f-4f07-aa02-26d91aaa3214

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
------------------	--------------

Submission d0fcd4ea-b3f8-4112-a47e-faec759104c8

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
------------------	--------------

Submission 8645fc4a-bece-44c1-80d3-b45f6f94ac09

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	

Submission 4f6607c4-6b85-41fd-a939-f33bae3d7b00

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

Submission ca2be4a6-e749-469d-b98a-93dc066e6acc					
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 reviewNot measured

Submission 54e94bfb-577b-4579-8ed8-eb88ad186676

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 reviewNot measured

Submission 8ec7d239-f799-438e-8fe9-1bb6394a49a7

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

Submission 6146c764-8fe6-4a86-8046-7778477f2ce7

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

Submission 3950adf9-665f-44d1-bb33-8e78bdb713d0

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

Submission f6899f85-7249-4408-bd32-80529c4def1f

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

Submission 4ce52860-fcb9-404f-867d-db6ffc0f94dc

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

Submission b9e9d7f7-5ce5-47b0-a67b-78da031148c0

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
------------------	--------------

Submission 8ae4a7c0-97cd-447a-9c58-7f4e39c0f014

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
------------------	--------------

Submission 0667cb37-cd47-49cf-93a1-47698ca729fc

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	

Submission 4012da48-6b0d-48ba-862a-9252d599d87b

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	

Submission e625725f-ecc0-4baf-b968-bdce899125d1

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 reviewNot measured

Submission fbad5bb3-f112-4f8a-bb9e-9def207856ff

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 reviewNot measured

Submission 1a963341-3836-4e59-b927-3e25494a79cd

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	10	7 / 24	0	0 / 0	0	0	0	0
cli:cli/claude-sonnet	10	5 / 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	10	0	none	0 / 10	0 / 10	0.0%	0
cli:cli/claude-sonnet	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 20 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/claude-sonnet, cli/codex passed the independent hidden tests		
Severity: PASS		Evidence: Model outcomes · Stage funnel · Appendix A
20 of 20 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		

Recommendations

Every rollout in this run passed; there is nothing to recommend beyond re-measuring on a harder exam.

Verification

RUN ID

live-20260907T090222Z

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

3596e3de9ba226bcd00ca746e26e31a8c0cd85b88f82b93d51ab27fd24af6877

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

SOURCE RECORD FILE

c67d426d290550b675391326f2941e25021b96275e8b3d0d01cd064c7189197a

sha256 of `fr-20260907-cd08adce.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

a609b424b13bd6faacc4379295cea88b0ec677bd3b285832daa135ec76fa8d11

sha256 of `fr-20260907-cd08adce.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

e40ef33940015815bdb26849b535417e759c1daadcb073fb9bb7802fbda7e6d7

sha256 of `fr-20260907-cd08adce.pdf`.

RUN EVIDENCE BUNDLE

cd08adce2a28569943c9ef42d4002456036f7ce669e26c4801160ddedfdafdac

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
b8eda0e7	cli/codex	0	submitted_passed	yes	Passed	pass	1	n/a	66.2s
1477727f	cli/codex	1	submitted_passed	yes	Passed	pass	1	n/a	57.2s
e59a52e2	cli/codex	2	submitted_passed	yes	Passed	pass	1	n/a	57.1s
d0fcd4ea	cli/codex	3	submitted_passed	yes	Passed	pass	1	n/a	46.2s
8645fc4a	cli/codex	4	submitted_passed	yes	Passed	pass	1	n/a	63.5s
4f6607c4	cli/codex	5	submitted_passed	yes	Passed	pass	1	n/a	60.5s
ca2be4a6	cli/codex	6	submitted_passed	yes	Passed	pass	1	n/a	57.8s
54e94bfb	cli/codex	7	submitted_passed	yes	Passed	pass	1	n/a	53.8s
8ec7d239	cli/codex	8	submitted_passed	yes	Passed	pass	1	n/a	56.8s
6146c764	cli/codex	9	submitted_passed	yes	Passed	pass	1	n/a	50.3s
3950adf9	cli/claude-sonnet	0	submitted_passed	yes	Passed	pass	1	n/a	24.2s
f6899f85	cli/claude-sonnet	1	submitted_passed	yes	Passed	pass	1	n/a	22.1s
4ce52860	cli/claude-sonnet	2	submitted_passed	yes	Passed	pass	1	n/a	17.0s
b9e9d7f7	cli/claude-sonnet	3	submitted_passed	yes	Passed	pass	1	n/a	21.9s
8ae4a7c0	cli/claude-sonnet	4	submitted_passed	yes	Passed	pass	1	n/a	21.4s
0667cb37	cli/claude-sonnet	5	submitted_passed	yes	Passed	pass	1	n/a	15.6s
4012da48	cli/claude-sonnet	6	submitted_passed	yes	Passed	pass	1	n/a	18.2s
e625725f	cli/claude-sonnet	7	submitted_passed	yes	Passed	pass	1	n/a	18.9s
fbad5bb3	cli/claude-sonnet	8	submitted_passed	yes	Passed	pass	1	n/a	16.4s
1a963341	cli/claude-sonnet	9	submitted_passed	yes	Passed	pass	1	n/a	26.3s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

cd08adce2a28569943c9ef42

d4002456036f7ce669e26c48

01160ddedfdafdac

RUN ID

live-

20260907T090222Z

STARTED

2026-09-

07T09:02:22.053916

+00:00

FINISHED

2026-09-

07T09:15:29.780105

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	4d50fc6ef69cc9adaa0033e397f56fc5d0eaeaa26cfac97ea2914ce3b96ed5b1
rollouts/0667cb37-cd47-49cf-93a1-47698ca729fc.json	2092dab814f297b26b3030470861b0da26ab9643f444e2ad412a50c08f6dca11
rollouts/1477727f-dbc3-450e-b2ed-e82f6d69d655.json	865a5df98ffa14b929a640031a62a287766dc7ab02ce83d43d77163552eb504f
rollouts/1a963341-3836-4e59-b927-3e25494a79cd.json	6845bef96afe967be4f21e26561b7dd7e2098486f53deac650d9aee968ca725e
rollouts/3950adf9-665f-44d1-bb33-8e78bdb713d0.json	f2a1936905cb4ebfeb283a0a72f31514d35563397d31bd4bf0406d323bbf954d
rollouts/4012da48-6b0d-48ba-862a-9252d599d87b.json	2cc7b2a5966189cca577594791cc04ca7305dd7a439b12be22c122d7735b94b9
rollouts/4ce52860-fcb9-404f-867d-db6ffcf0f94dc.json	7fecf3a9e9897a2bcb3eb38ccaf3ed5ce9bee012e7539633c103e055c59b442f
rollouts/4f6607c4-6b85-41fd-a939-f33bae3d7b00.json	1d39c61936b8f9069704f1ba4e6090ea68e650702229f4b80e78b0bfda3bafef7
rollouts/54e94bfb-577b-4579-8ed8-eb88ad186676.json	e7334209c42ddd8846b0c2a43b63edc01629896517f0f78a6faea2a942679ede
rollouts/6146c764-8fe6-4a86-8046-7778477f2ce7.json	3fffd3da00db0b55e6d337804e6668c87d5d662ac30054d3e32be362550fe921f
rollouts/8645fc4a-bece-44c1-80d3-b45f6f94ac09.json	4324ad56d93f96cb414eebd24d1af5ed07689a75e46599adb8accbfff83b4f2
rollouts/8ae4a7c0-97cd-447a-9c58-7f4e39c0f014.json	b2c1daa80dbfc283e2a682566f0212edb97d1260e0ff1b0cddb5d044d5f08caa
rollouts/8ec7d239-f799-438e-8fe9-1bb6394a49a7.json	813adf4fedba739e90830d1cd70a3c07853adba06e3b9fd429b7080df93a487a
rollouts/b8eda0e7-467a-47b0-a562-e998a3c58062.json	287b27a0a1eebe22c403cb16d23ce4908c0d06a09e9e6aece7f2a230d1a596b
rollouts/b9e9d7f7-5ce5-47b0-a67b-78da031148c0.json	7326e64fa8e27c12b9acf57ac531a9d4a8255caeb40f3d839735a5c93a4f2f57
rollouts/ca2be4a6-e749-469d-b98a-93dc066e6acc.json	9f268e0e3613f786b65bcf627de81000ddb0049798a036d5be066827ecbf3069
rollouts/d0fcd4ea-b3f8-4112-a47e-faec759104c8.json	46b356221f8b1f9cdc8e183e04f64797183d9f400831d99dd3312ae1ba68891b
rollouts/e59a52e2-a69f-4f07-aa02-26d91aaa3214.json	18b36e3eeff4e889e7e65e459f7392605cdb713b5eca1a033538d98594c74be9
rollouts/e625725f-ecc0-4baf-b968-bdce899125d1.json	c00eca8107b6fedf0b5f8c76b0bccb2cb6ac1265fe625da2c2461456f57fcc76
rollouts/f6899f85-7249-4408-bd32-80529c4def1f.json	a695ece0af7783b68fd4e1aefd2beea82cb479e275bed5fa1cd7211bb6dd868c
rollouts/fbad5bb3-f112-4f8a-bb9e-9def207856ff.json	b7a61de5b09fb53f44728d12845f179865b1865e1eefa37f4fb0aeffc5f7730
spec	26fc48c29ae61b08841c8f2b906ca64cdf05650e26452f759f052c5d995e78b
summary	4c9a2c9cf1c77ac1fd4a255f2a2b2390fec187ca3712ba3a5a0323e976218346

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.

b8eda0e7 · cli:cli/codex · 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	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
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.

1477727f · cli:cli/codex · seed 1 · 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	160
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.

e59a52e2 · cli:cli/codex · seed 2 · 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	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

d0fcd4ea · cli:cli/codex · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	read_file	assets/records.json	ok	295
3	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

8645fc4a · cli:cli/codex · seed 4 · 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	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (41 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.

4f6607c4 · cli:cli/codex · seed 5 · 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	160
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.

ca2be4a6 · cli:cli/codex · seed 6 · 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	160
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.

54e94bfb · cli:cli/codex · seed 7 · 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

STEP	TOOL	PATH	RESULT	CHARS
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
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.

8ec7d239 · cli:cli/codex · seed 8 · 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	160
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.

6146c764 · cli:cli/codex · seed 9 · 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	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

3950adf9 · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520

STEP	TOOL	PATH	RESULT	CHARS
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

f6899f85 · cli:cli/claude-sonnet · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

4ce52860 · cli:cli/claude-sonnet · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

b9e9d7f7 · cli:cli/claude-sonnet · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

8ae4a7c0 · cli:cli/claude-sonnet · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

0667cb37 · cli:cli/claude-sonnet · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

4012da48 · cli:cli/claude-sonnet · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

e625725f · cli:cli/claude-sonnet · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290

STEP	TOOL	PATH	RESULT	CHARS
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

fbad5bb3 · cli:cli/claude-sonnet · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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.

1a963341 · cli:cli/claude-sonnet · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	annotation-task.json	ok	3 290
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 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 1 of 1 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-8f0b9493 · record digest 61073e4730b6a0ec... · harness clean · containment clean

Executive summary

MODELS TESTED 2	CERTIFIED PASSES 20 of 20 graded rollouts	SUBMITTED 20 of 20 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 20 attempts
---------------------------------------	--	--	--

Purpose

This report presents a controlled evaluation of 2 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

20 of 20

graded rollouts passed the independent hidden tests and were certified

20 of 20 graded rollouts passed. cli/claude-sonnet, cli/codex completed the applicable annotation workflow and were accepted by the independent annotation grader.

Key findings

F-01 cli/claude-sonnet, cli/codex passed the independent hidden tests

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: 827.6s.

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 14520s of wall clock, at 10 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	20 (10 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-20260907-8f0b9493
ISSUED	2026-09-07
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)
2026-09-07 08:48:26	Evaluation run started
2026-09-07 09:02:14	Evaluation run finished
2026-09-07 09:02:14	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Claude Code CLI (Sonnet, subscription)	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 docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED 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: 20. Graded (evaluable) rollouts: 20. 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
Codex CLI (GPT, subscription) cli/codex	10	10	10 / 10	0	0	10 / 10	55.2s	Passed
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	10	10	10 / 10	0	0	10 / 10	26.1s	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	PASS@3	PASS@5	PASS@10
cli/codex	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
cli/claude-sonnet	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1, 3, 5, 10; the sealed record carries every k up to n.

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/codex	100.0%	[72%, 100%]	n/a	n/a	55.2s	0.0%	none
cli/claude-sonnet	100.0%	[72%, 100%]	n/a	n/a	26.1s	0.0%	none

Tokens and cost read n/a where a lane reports no telemetry (a flat-rate CLI) — an absent measurement, not zero. pass@1 and its interval are over evaluable rollouts only.

This is not a leaderboard. Not separated (intervals overlap): cli/codex vs cli/claude-sonnet. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	CLI:CLI/CODEX	CLI:CLI/CLAUDE-SONNET
Inspect	10 / 10	10 / 10
Annotate	10 / 10	10 / 10
Validate	10 / 10	10 / 10
Submit	10 / 10	10 / 10
Graded	10 / 10	10 / 10
Passed	10 / 10	10 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/codex	Passed	Passed	10 / 10	0
cli:cli/claude-sonnet	Passed	Passed	10 / 10	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 6562b25a-1eda-40c0-a03f-9e1f8ba8f0af

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
------------------	--------------

Submission 1ac06c7b-b00e-424c-b953-ff4eb6624879

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
------------------	--------------

Submission bc443e0e-e6ed-4a27-921c-329f22a985fb

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
------------------	--------------

Submission 8506c33c-683b-4b15-9d42-ccb7e48f647f

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
------------------	--------------

Submission 4f40821b-cd6e-463a-a147-ade0a9f363

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

Submission 26f398a6-2b2f-4472-9f97-92b5f39556d1

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

Submission db5157cc-dd4a-4d52-9ee6-1d092d7fd934

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 reviewNot measured

Submission 1ba34a8d-93bc-4bc1-8192-75375a5fab9e

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 reviewNot measured

Submission afea3a7c-3add-4776-b178-b17336e18472

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

Submission 1e91b362-e78f-414e-9623-3fb6631635a1

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

Submission c7c0b5b1-593a-454b-8b33-a5640dc80522

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		
Submission 1956f3b9-0fb8-44b0-864f-43a3710414b0					
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		
Submission 5540d66f-90f0-44f4-a346-9a4b15dcc7bf					
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		

Submission f75f7ffd-0c16-420a-9b5c-eb061c52de51

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
------------------	--------------

Submission fc30d37d-23e9-43ce-ac04-874bc7b40d89

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
------------------	--------------

Submission 7cf58734-0777-469e-943f-6754f9df87d6

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	

Submission cafe5c18-54b9-4474-9744-04aa14532fab

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	

Submission 272dafb0-9d9d-43cd-8631-520dbcb64663

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 reviewNot measured

Submission 8052afd0-5e12-46f8-8721-696df6b94426

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 reviewNot measured

Submission a5be9e58-491f-47fc-a268-b5bfec370d0a

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	10	7 / 24	0	0 / 0	0	0	0	0
cli:cli/claude-sonnet	10	5 / 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	10	0	none	0 / 10	0 / 10	0.0%	0
cli:cli/claude-sonnet	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 20 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/claude-sonnet, cli/codex passed the independent hidden tests		
Severity: PASS		Evidence: Model outcomes · Stage funnel · Appendix A
20 of 20 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		

Recommendations

Every rollout in this run passed; there is nothing to recommend beyond re-measuring on a harder exam.

Verification

RUN ID

live-20260907T084826Z

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

61073e4730b6a0ec35b5bcd8e984e7382e6646a911b99bcfa88c145d7756e8fa

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

SOURCE RECORD FILE

7e1edb3e0a187b4c0dadd8214949203d2c4f540944f94e4e4633375dd2a07b35

sha256 of `fr-20260907-8f0b9493.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

e8ae6e8b4f97c922fda03428ba4d00d3ef36aa242ddbafc8576c3a7ecb1c69ea

sha256 of `fr-20260907-8f0b9493.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

b45e241c24b5f1fda385cca59e7ea72c6f3a1ced2f671d7aee0b0f76ebf1dba4

sha256 of `fr-20260907-8f0b9493.pdf`.

RUN EVIDENCE BUNDLE

8f0b94935fb71c5af554bddd6c2332990bbdfeee1b5f2275b97e4f490a04c9da

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
6562b25a	cli/codex	0	submitted_passed	yes	Passed	pass	1	n/a	57.5s
1ac06c7b	cli/codex	1	submitted_passed	yes	Passed	pass	1	n/a	68.6s
bc443e0e	cli/codex	2	submitted_passed	yes	Passed	pass	1	n/a	55.5s
8506c33c	cli/codex	3	submitted_passed	yes	Passed	pass	1	n/a	58.9s
4f40821b	cli/codex	4	submitted_passed	yes	Passed	pass	1	n/a	44.9s
26f398a6	cli/codex	5	submitted_passed	yes	Passed	pass	1	n/a	45.2s
db5157cc	cli/codex	6	submitted_passed	yes	Passed	pass	1	n/a	46.6s
1ba34a8d	cli/codex	7	submitted_passed	yes	Passed	pass	1	n/a	60.8s
afea3a7c	cli/codex	8	submitted_passed	yes	Passed	pass	1	n/a	61.5s
1e91b362	cli/codex	9	submitted_passed	yes	Passed	pass	1	n/a	52.6s
c7c0b5b1	cli/claude-sonnet	0	submitted_passed	yes	Passed	pass	1	n/a	22.1s
1956f3b9	cli/claude-sonnet	1	submitted_passed	yes	Passed	pass	1	n/a	26.3s
5540d66f	cli/claude-sonnet	2	submitted_passed	yes	Passed	pass	1	n/a	24.0s
f75f7ffd	cli/claude-sonnet	3	submitted_passed	yes	Passed	pass	1	n/a	22.9s
fc30d37d	cli/claude-sonnet	4	submitted_passed	yes	Passed	pass	1	n/a	29.0s
7cf58734	cli/claude-sonnet	5	submitted_passed	yes	Passed	pass	1	n/a	20.1s
cafe5c18	cli/claude-sonnet	6	submitted_passed	yes	Passed	pass	1	n/a	34.7s
272dafb0	cli/claude-sonnet	7	submitted_passed	yes	Passed	pass	1	n/a	20.1s
8052afd0	cli/claude-sonnet	8	submitted_passed	yes	Passed	pass	1	n/a	38.4s
a5be9e58	cli/claude-sonnet	9	submitted_passed	yes	Passed	pass	1	n/a	23.4s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

8f0b94935fb71c5af554bddd

6c2332990bbdfeee1b5f2275

b97e4f490a04c9da

RUN ID

live-

20260907T084826Z

STARTED

2026-09-

07T08:48:26.572120

+00:00

FINISHED

2026-09-

07T09:02:14.183999

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	e6c9b5eb366aa21d11f6cbee5cd0815fa1ff132f0aec3375a7f4c496a5841521
rollouts/1956f3b9-0fb8-44b0-864f-43a3710414b0.json	a7d2880c3191fe4c43ba416ee9dc6e027562691680da70061087a1b2ada6ae14
rollouts/1ac06c7b-b00e-424c-b953-ff4eb6624879.json	08d4173073f7a28c467ce25c54acdbc4d6340581ac2c7cd4b6562f7e54061831
rollouts/1ba34a8d-93bc-4bc1-8192-75375a5fab9e.json	dca606bea0745c91d966efb8e9f07270e07e7ce0c878b1b20ea22155cb5b6128
rollouts/1e91b362-e78f-414e-9623-3fb6631635a1.json	88a6ef6c4144931c13f50a08495972059bce3eba0ddf8247aa699f226231c3de
rollouts/26f398a6-2b2f-4472-9f97-92b5f39556d1.json	261e5a7dbef802a2e1ea18a27fc48b74c02f2ec63ab034cfa0e671057e996983
rollouts/272dafb0-9d9d-43cd-8631-520dbcb64663.json	21016aa58d232a44b460fb0afde3240a400cceba4807017579d53a268c16e37dc
rollouts/4f40821b-cd6e-463a-a147-ade0a9f363.json	08d529ff305b17544bd425196141d4b0413e923fdaed64573acced7e98bfa1b6
rollouts/5540d66f-90f0-44f4-a346-9a4b15dcc7bf.json	fa4f452cef6c28c17ffb8cb60e2e93c13d85eb9744e432bebc4bf8fbb6d7a6bc
rollouts/6562b25a-1eda-40c0-a03f-9e1f8ba8f0af.json	5bb5c4fefe605934cd5955d8bf3a312564586025d73414c936d9e2eb0b76bfd4
rollouts/7cf58734-0777-469e-943f-6754f9df87d6.json	b44c596aec0c7f500328d8cca02667909a339d976a5c85270f69af899835939a
rollouts/8052afd0-5e12-46f8-8721-696df6b94426.json	26ff4e7b0440454950d391189e98b7943fa5691d2b899d3c2e9b83bf35e6c618
rollouts/8506c33c-683b-4b15-9d42-ccb7e48f647f.json	4216f197805c890501bb14a1a4908612cc8f87f1c4d99d68cc2e03033a7848b1
rollouts/a5be9e58-491f-47fc-a268-b5bfec370d0a.json	61fc84cbc5252cee0c3fb491ca6db65dde6fb7ed704666901635290b18106edf
rollouts/afea3a7c-3add-4776-b178-b17336e18472.json	95ad233c81673df2209b0cb12000ab7d529bb831ada55648e4887714fe72bee5
rollouts/bc443e0e-e6ed-4a27-921c-329f22a985fb.json	c5673fd4a59a46e624c8a30e471a67907aa9885fcd306cd07d73f414be659fb7
rollouts/c7c0b5b1-593a-454b-8b33-a5640dc80522.json	262062126a8185eebe8ad8009d9ebc9b09678c06ff0963e4c4d0e03d226f72cf
rollouts/cafe5c18-54b9-4474-9744-04aa14532fab.json	955f877e94c47490eae8c70f4d1d95806bb44e10ba9070fdd5bc3c101415c0b6
rollouts/db5157cc-dd4a-4d52-9ee6-1d092d7fd934.json	f0710dce9305b7cbdf36e3d47bb4625294c963cb4b32df8ba04c9146351467ca
rollouts/f75f7ffd-0c16-420a-9b5c-eb061c52de51.json	9a8c4f3f24835c636c23283a5ab41420509717f21073f3cea583b3ab8e3fb134
rollouts/fc30d37d-23e9-43ce-ac04-874bc7b40d89.json	86a3b74661d28b3d262ee2434e84c28fb8796138b2d1fd42a4a47f8df7188096
spec	1b78e911fd84bed7461339a27fab18bd6e6fd4af0de55461facb4de33b134f17
summary	b38dc46ddf3a8de87a818ebf55223d8a9d9174c026c0f5afc41460e6543b33bd

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.

6562b25a · cli:cli/codex · 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	160
7	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.

1ac06c7b · cli:cli/codex · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	list_files	.	ok	80
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	160
7	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.

bc443e0e · cli:cli/codex · seed 2 · 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	160
7	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.

8506c33c · cli:cli/codex · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	list_files	assets	ok	18
3	read_file	assets/message.txt	ok	69
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
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.

4f40821b · cli:cli/codex · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	read_file	assets/message.txt	ok	69
3	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

26f398a6 · cli:cli/codex · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	read_file	assets/message.txt	ok	69
3	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

db5157cc · cli:cli/codex · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	read_file	assets/message.txt	ok	69
3	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

1ba34a8d · cli:cli/codex · seed 7 · 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	160
7	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.

afea3a7c · cli:cli/codex · seed 8 · 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	160
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.

1e91b362 · cli:cli/codex · seed 9 · 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	160
7	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.

c7c0b5b1 · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

1956f3b9 · cli:cli/claude-sonnet · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

5540d66f · cli:cli/claude-sonnet · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	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.

f75f7ffd · cli:cli/claude-sonnet · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	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.

fc30d37d · cli:cli/claude-sonnet · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64

STEP	TOOL	PATH	RESULT	CHARS
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

7cf58734 · cli:cli/claude-sonnet · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

cafe5c18 · cli:cli/claude-sonnet · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	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.

272dafb0 · cli:cli/claude-sonnet · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	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.

8052afd0 · cli:cli/claude-sonnet · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 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.

a5be9e58 · cli:cli/claude-sonnet · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	annotation-task.json	ok	2 826
2	write_file	submission.json	ok	64
3	run_tests	—	ok	160
4	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.

11 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.structured-data-1` → `fr-20260907-cd08adce` · `recordDigest` `3596e3de9ba226bcd00ca746e26e31a8c0cd85b88f82b93d51ab27fd24af6877`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-8f0b9493` · `recordDigest` `61073e4730b6a0ec35b5bcd8e984e7382e6646a911b99bcfa88c145d7756e8fa`

12 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/. Third-party open source reproduced in an exam keeps its own licence.