VVDEX FORGE

CERTIFIED EVALUATION CAMPAIGN

Gemini and OpenCode pilots — image, audio, text and structured, one attempt each

4 fixed tasks · 10 sealed attempt records · 6 graded attempts · 2 certified passes

CURRENT

CAMPAIGN REPORT

2026-09-07

Evaluation scope:

gemini-3.6-flash

cli/opencode-muse-spark

Prepared by:

VVDex Forge — regenerated from verified sealed records only

`fc-6f40beb6605e` · 10 record(s) · every value recomputable

Contents

01 Executive summary

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

FIXED TASKS	SEALED RECORDS	MODEL LANES	ATTEMPTS
4	10	2	10

GRADED ATTEMPTS	CERTIFIED PASSES	LANE ERRORS (NOT GRADED)	WITHHELD / INVALID
6	2	4	3

Purpose

This report presents a controlled evaluation of 2 model lane(s) against 4 fixed certified annotation tasks, between 1 and 3 sealed attempt records per 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 between 1 and 3 sealed attempt records per task (10 sealed attempts in total). The sealed tasks cover audio, image, structured, text inputs. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

2

certified pass(es) across 6 graded attempts (10 attempts, 4 lane error(s) not graded) in 7 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-6f40beb6605e · generated from 10 verified sealed record(s)

Key findings across the campaign

- `audio-events-1` · `gemini-3.6-flash` · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `image-object-1` · `gemini-3.6-flash` · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `image-object-1` · `cli/opencode-muse-spark` · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 1 excluded (0 lane errors, 1 withheld, 0 invalid).
- `structured-data-1` · `gemini-3.6-flash` · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `structured-data-1` · `cli/opencode-muse-spark` · **ATTENTION** — 1 certified passes / 1 graded attempts; 0 model failures; 1 excluded (0 lane errors, 1 withheld, 0 invalid).
- `text-labeling-1` · `gemini-3.6-flash` · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `text-labeling-1` · `cli/opencode-muse-spark` · **ATTENTION** — 1 certified passes / 1 graded attempts; 0 model failures; 1 excluded (0 lane errors, 1 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION · AUDIO-EVENTS-1 VVDEX.ANNOTATION.AUDIO-EVENTS-1	ANNOTATION · IMAGE-OBJECT-1 VVDEX.ANNOTATION.IMAGE-OBJECT-1	ANNOTATION · STRUCTURED-DATA-1 VVDEX.ANNOTATION.STRUCTURED-DATA-1	ANNOTATION · TEXT-LABELING-1 VVDEX.ANNOTATION.TEXT-LABELING-1
cli/opencode-muse-spark	—	0/1 + 1 withheld	1/1 + 1 withheld	1/1 + 1 withheld
gemini-3.6-flash	0/0 + 1 lane error	0/0 + 1 lane error	0/0 + 1 lane error	0/0 + 1 lane error

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	`vvdex.annotation.audio-events-1`	Constructed non-speech audio is reviewed for events, temporal segments, silence and category consistency.
Multimodal annotation	`vvdex.annotation.image-object-1`	A constructed image is reviewed for classification, object boxes, attributes and keypoints.
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 four exams

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
cli/opencode-muse-spark	**2**	3	0	3	0
gemini-3.6-flash	**0**	0	4	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

4 of 4 exams include harness-suspect records (audio-events-1, image-object-1, structured-data-1, text-labeling-1). Each fired rule is classified below; a verdict-bearing rule the record cannot clear from its own evidence withholds the affected cells. **0 cell(s) withheld** on that basis.

`audio-events-1 · R1 · diagnostic`

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

`image-object-1 · R1 · diagnostic`

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

`image-object-1 · R4 · row-level`

containment verdicts are already enforced on every rollout; the run-level flag adds no further adjustment.

`structured-data-1 · R4 · row-level`

containment verdicts are already enforced on every rollout; the run-level flag adds no further adjustment.

`structured-data-1 · R1 · diagnostic`

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

`text-labeling-1 · R1 · diagnostic`

an API failure is a per-rollout lane event, already recorded on that rollout as lane_error; it cannot alter another rollout's grading. It lowers coverage, not verdicts.

`text-labeling-1 · R4 · row-level`

containment verdicts are already enforced on every rollout; the run-level flag adds no further adjustment.

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.

EXAM	LANE	SEED	EXCLUSION	RECORDED REASON
audio-events-1	gemini-3.6-flash	0	lane error	HTTP 429
image-object-1	cli/opencode-muse-spark	0	withheld	withheld (conditions unproven)
image-object-1	gemini-3.6-flash	0	lane error	HTTP 429
structured-data-1	cli/opencode-muse-spark	0	withheld	withheld (conditions unproven)
structured-data-1	gemini-3.6-flash	0	lane error	HTTP 429
text-labeling-1	cli/opencode-muse-spark	0	withheld	withheld (conditions unproven)
text-labeling-1	gemini-3.6-flash	0	lane error	HTTP 429

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
audio-events-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.
image-object-1	answer_key	**Applied consistently to 3 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 3 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 3 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 · audio-events-1

`vvdex.annotation.audio-events-1 · record fr-20260907-67c07cce · record digest 1f595aa9846136ff… · harness SUSPECT · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 0 graded rollouts	0 of 0 graded rollouts that called submit	1 of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam `vvdex.annotation.audio-events-1`. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 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 `model_api_failure` — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	model_api_failure × 1 (gemini-3.6-flash)
F-03	The harness suspected itself during this run
F-04	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: 1.2s.

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 2280s 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.audio-events-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.audio-events-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

`3792d5564b24693e88dc18df97eb13e485bbb63ecc5950209ab92601e06600ae`

EXAM FINGERPRINT

`7e749a2ffb885355e770a090…`

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

Evaluation configuration

REPORT	fr-20260907-67c07cce
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.audio-events-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 08:45:27	Evaluation run started
2026-09-07 08:45:28	Evaluation run finished
2026-09-07 08:45:28	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.6-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	2280s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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: 0. Not graded — lane errors: 1. 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
gemini-3.6-flash `gemini-3.6-flash`	1	0	not graded	0	1	–	–	nothing

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
gemini-3.6-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; 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	GEMINI:GEMINI-3.6-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.6-flash	nothing	nothing	0 / 0	1

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

Annotation evidence

Modality: audio. **Rubric:** 1.0.

Constructed non-speech audio is reviewed for events, temporal segments, silence and category consistency.

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
gemini:gemini-3.6-flash	1	0 / 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
gemini:gemini-3.6-flash	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.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.6-flash	0	n/a	n/a	n/a — telemetry unavailable	0 / 0	0.0%

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 LANE / INFRA

model_api_failure × 1 (gemini-3.6-flash)

Severity: LANE / INFRA | Evidence: Appendix A · Appendix C

The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.

F-02 · derived from the recorded run data

F-03 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-03 · derived from the recorded run data

F-04 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-04 · 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

The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES

gemini-3.6-flash

WHAT WE WOULD LOOK AT FIRST

The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R–02 **For anyone reading the percentages**

Commission a run at n ≥ 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-20260907T084527Z

EXAM FINGERPRINT

7e749a2ffb885355e770a09007a7b2465edf1d8a50768441b8872ff9f951ed5e

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

1f595aa9846136ff6fcd075c0a6f491a969633f119bb0c63505f70a92e0e6a02

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

SOURCE RECORD FILE

9117d15d9f8a7519d353ffbb4fa33a34e3f6e2d5993d12de792f00d40201983a

sha256 of fr-20260907-67c07cce.json as written beside the standalone report.

SOURCE STANDALONE REPORT

3d3ad1c8354e77b08fd3ceaf9529a8f00e5cdab74cd4087b3bb4ba84fc1b1416

sha256 of fr-20260907-67c07cce.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

ffe67ac32b2b5eb79514687c82ce8045f89ba92413731eaed81c511ef826d96b

sha256 of fr-20260907-67c07cce.pdf.

RUN EVIDENCE BUNDLE

67c07cce2ed77ecfeb3ebef2ef24f1107f11a817c0b7ba8d1cefa2c3cde8f28f

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

CERTIFIED EVIDENCE SEAL

3792d5564b24693e88dc18df97eb13e485bbb63ecc5950209ab92601e06600ae

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
e78fec44	gemini-3.6-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	891ms

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
67c07cce2ed77ecfeb3ebef2ef24f1107f11a817c0b7ba8d1cefa2c3cde8f28f	live-20260907T084527Z	2026-09-07T08:45:27.534846+00:00	2026-09-07T08:45:28.767445+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	84af6f50ede9800c3723837aed4ec004d7d7df8c23a0b4901bbec2bc9247572f
rollouts/e78fec44-66e6-4b28-bdcc-38963d93efbd.json	7526ac2d224a5418fd76b7a529a1519ab46f602a86e3dead9eddc27da5b7a74d
spec	07ce99796d9e158d658c5588514f15f1f136ca91c28e7e2934d335f5aac29d9a
summary	7cc01b31e3656c7e0d59546b3bedf7015fb2807e92dfb464eb7337fb944feb8b

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.

e78fec44 · gemini:gemini-3.6-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
No tool call was recorded for this rollout.				

Submission diff withheld by the public-disclosure policy. 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 3 · image-object-1

`vvdex.annotation.image-object-1 · record fr-20260907-54b8ff2d · record digest 020983794ee70906… · harness SUSPECT · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 0 graded rollouts	0 of 0 graded rollouts that called submit	1 of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam `vvdex.annotation.image-object-1`. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 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 `model_api_failure` — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	model_api_failure × 1 (gemini-3.6-flash)
F-03	The harness suspected itself during this run
F-04	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: 658ms.

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 2280s 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-20260907-54b8ff2d
ISSUED	2026-09-07
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)
2026-09-07 09:03:09	Evaluation run started
2026-09-07 09:03:09	Evaluation run finished
2026-09-07 09:03:09	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.6-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	2280s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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: 0. Not graded — lane errors: 1. 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
gemini-3.6-flash `gemini-3.6-flash`	1	0	not graded	0	1	–	–	nothing

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
gemini-3.6-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; 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	GEMINI:GEMINI-3.6-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.6-flash	nothing	nothing	0 / 0	1

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.

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
gemini:gemini-3.6-flash	1	0 / 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
gemini:gemini-3.6-flash	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.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.6-flash	0	n/a	n/a	n/a — telemetry unavailable	0 / 0	0.0%

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 LANE / INFRA

model_api_failure × 1 (gemini-3.6-flash)

Severity: LANE / INFRA | Evidence: Appendix A · Appendix C

The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.

F-02 · derived from the recorded run data

F-03 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-03 · derived from the recorded run data

F-04 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-04 · 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

The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES

gemini-3.6-flash

WHAT WE WOULD LOOK AT FIRST

The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R–02 **For anyone reading the percentages**

Commission a run at n ≥ 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-20260907T090309Z

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

020983794ee70906ee3c9162e3326cb5575cbf75cd4963d811d1378a966c732c

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

SOURCE RECORD FILE

3e8c8ed06064f930e935b895e9a1e7baae082b526da3f5527ac724cad97bf59c

sha256 of fr-20260907-54b8ff2d.json as written beside the standalone report.

SOURCE STANDALONE REPORT

ee33fb7d8799310a84d6e160e301a3949f9061aa81c8510f0582220947b557e6

sha256 of fr-20260907-54b8ff2d.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

24d0ec19f0da7bd24c2fac3d0be4fff80e4f745cd0c70312a03fe24c78ae5d78

sha256 of fr-20260907-54b8ff2d.pdf.

RUN EVIDENCE BUNDLE

54b8ff2d8e72becf9d8a24a48e89f8b70cc4175f8d4cb370791d7db68c0a0fa7

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-20260907-9bee4b48`. 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
32567d64	gemini-3.6-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	379ms

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
54b8ff2d8e72becf9d8a24a48e89f8b70cc4175f8d4cb370791d7db68c0a0fa7	live-20260907T090309Z	2026-09-07T09:03:09.116296+00:00	2026-09-07T09:03:09.774792+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	eec85611d0e75219a301653d4b9f33f63b1aab426e01af785895e57619c03d14
rollouts/32567d64-111d-4fae-9c92-9ae7f144e920.json	f17678ab28f9fc80482ca89492cc8b6d8ae940895c2de420314ad874ca88f665
spec	3f72edfe682b56235700619d2163afbe70129a0800784af748882d72b0a294b7
summary	aaafc4311d585bec679625e88223bde1d1ebc6a1434fca0bc27fe7e6da1bac0a

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.

32567d64 · gemini:gemini-3.6-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
No tool call was recorded for this rollout.				

Submission diff withheld by the public-disclosure policy. 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 2 of 3 · image-object-1

`vvdex.annotation.image-object-1 · record fr-20260907-9bee4b48 · record digest b08ceac9eef6a1d0… · harness SUSPECT · {"breached": 0, "unverified": 1}`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 1 graded rollouts	1 of 1 graded rollouts that called submit	0 of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vv`dex.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/opencode-muse-spark)
F-03	1 rollout(s) could not be proven contained
F-04	The harness suspected itself during this run
F-05	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: 166.3s.

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 14520s 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-20260907-9bee4b48
ISSUED	2026-09-07
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)
2026-09-07 08:45:27	Evaluation run started
2026-09-07 08:48:13	Evaluation run finished
2026-09-07 08:48:13	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	0 / 1	1	0	1 / 1	166.0s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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 6e0dd6df-38ce-4cd1-84c5-7700eb5ecfc5

Score	Items	Valid / invalid	Types
0.7665	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.5433
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.0206	keypointPrecision	1
keypointRecall	1	matchedCount	4	mismatchCount	3

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	0.5433	–	–	
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)

0.7665

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/opencode-muse-spark	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/opencode-muse-spark	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/opencode-muse-spark)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted annotations rejected by the independent grader.

F-02 · derived from the recorded run data

F-03 ATTENTION

1 rollout(s) could not be proven contained

Severity: ATTENTION | Evidence: Appendix A

No evidence establishes that the taker stayed inside the exam channel, and none establishes that it left. These rows are excluded from capability arithmetic and are neither passes nor model failures.

F-03 · derived from the recorded run data

F-04 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R4 containment unproven in 1/1 rollouts (lanes ['cli/opencode-muse-spark']) — no evidence the taker was confined to the exam channel. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-04 · derived from the recorded run data

F-05 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-05 · 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/opencode-muse-spark

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 ≥ 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-20260907T084527Z

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

b08ceac9eef6a1d09760e410f2b09a12c6de4c221c95cdc8aef20a6fc5ef776f

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

SOURCE RECORD FILE

9fb97a6020125bd1b85192bc3075aee1764280434a811955657db4ceae869200

sha256 of fr-20260907-9bee4b48.json as written beside the standalone report.

SOURCE STANDALONE REPORT

fa7bb48bb1a6b2db8557e58580d4e343852cfa909b321250ef467db70e72bba9

sha256 of fr-20260907-9bee4b48.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

538afc5f3b2ba1ed9f3d473074af1b767eab62c884fcd971de5b9969e8ba48d6

sha256 of fr-20260907-9bee4b48.pdf.

RUN EVIDENCE BUNDLE

9bee4b48630da0da0f284fbd2fd6fd4a91b1edb5605a08bdb58e1ecbc53fa451

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
6e0dd6df	cli/opencode-muse-spark	0	submitted_failed	yes	Graded	fail	1	37 809	166.0s

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	RUN ID	STARTED	FINISHED
9bee4b48630da0da0f284fbd2fd6fd4a91b1edb5605a08bdb58e1ecbc53fa451	live-20260907T084527Z	2026-09-07T08:45:27.439994+00:00	2026-09-07T08:48:13.745146+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	dd4f59877065786ba8b63238fd1580c82fc23f5effbb04381a24b04a376be6a4
rollouts/6e0dd6df-38ce-4cd1-84c5-7700eb5ecfc5.json	21b7e5a076b3ff8ae9ef31fb5836c4732019246a2b9fc9aec924a5ab7b7dac86
spec	2e55ba653db9c66f1259f8ed3a89572f92fd366a4571cdd877c808860590061b
summary	e055d8065d9e5b6ea3fc2f1c6ff1be06a0c6597ed083bc8ee0e9e300431421f8

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.

6e0dd6df · cli:cli/opencode-muse-spark · 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	submission.json	ok	74
4	read_file	assets/objects.png	ok	397
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	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.

12 Attempt record 3 of 3 · image-object-1

`vvdex.annotation.image-object-1 · record fr-20260907-e2134530 · record digest 80e72cacb9087660… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 1 graded rollouts	1 of 1 graded rollouts that called submit	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/opencode-muse-spark)
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: 71.0s.

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 14520s 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-20260907-e2134530
ISSUED	2026-09-07
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)
2026-09-07 09:22:03	Evaluation run started
2026-09-07 09:23:13	Evaluation run finished
2026-09-07 09:23:13	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	0 / 1	1	0	1 / 1	70.7s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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 ec02b5a5-c4c8-4d27-9b72-a23926affb18

Score	Items	Valid / invalid	Types
0.7296	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0	boxIoU	0.4793
boxPrecision	0	boxRecall	0	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.04	keypointPrecision	1
keypointRecall	1	matchedCount	4	mismatchCount	3

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	—	—	
boxRecall	recorded	N/A	0	—	—	
boxF1	recorded	N/A	0	—	—	
boxIoU	recorded	N/A	0.4793	—	—	
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)

0.7296

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/opencode-muse-spark	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/opencode-muse-spark	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/opencode-muse-spark)

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/opencode-muse-spark

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 ≥ 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-20260907T092203Z

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

80e72cacb9087660e9efeab8e1ab8c7635a7463c8d2b7be1d7900839ea280dd8

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

SOURCE RECORD FILE

e1cd84e2e14ccccdf9cd49dfa216bccaf566cf43d4d395ac0f315dfbea62c753

sha256 of fr-20260907-e2134530.json as written beside the standalone report.

SOURCE STANDALONE REPORT

f568f35ee457e8a346b351217d1f9e2e51f52bd87ca41de9c76d6324c22aa405

sha256 of fr-20260907-e2134530.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

d0f969b83f39e7875a07ecaf7102dc3665351bd9256b10e44860062c4e60abe0

sha256 of fr-20260907-e2134530.pdf.

RUN EVIDENCE BUNDLE

e21345300d3e1c496e421d19e7e9a2f4c1db8c1a01816c3477804c68f84b8481

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-20260907-54b8ff2d`. 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
ec02b5a5	cli/opencode-muse-spark	0	submitted_failed	yes	Graded	fail	1	37 905	70.7s

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	RUN ID	STARTED	FINISHED
e21345300d3e1c496e421d19e7e9a2f4c1db8c1a01816c3477804c68f84b8481	live-20260907T092203Z	2026-09-07T09:22:03.019052+00:00	2026-09-07T09:23:13.984211+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	198f949c10fd44f71a140ade5fca0e595d8b26b15b930a3de9105e0257df9690
rollouts/ec02b5a5-c4c8-4d27-9b72-a23926affb18.json	1c9b53cdd1265b34beaebf524e9581e1568758d10a75fb729b99ca9e3f436470
spec	574e2640ff4ca349381c94f98221d4602fa6980253de52db311234e12cdd1d45
summary	41249d492d206bca6f9c7ed2fbc6506e8f48c41d6f42213c75dd86bd56386314

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.

ec02b5a5 · cli:cli/opencode-muse-spark · 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	160
7	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.

13 Attempt record 1 of 3 · structured-data-1

`vvdex.annotation.structured-data-1 · record fr-20260907-30aad916 · record digest 9bcf06ca2f1d923f… · harness SUSPECT · {"breached": 0, "unverified": 1}`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 1 graded rollouts	1 of 1 graded rollouts that called submit	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

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 `unverified_isolation` — .

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	unverified_isolation × 1 (cli/opencode-muse-spark)
F-03	1 rollout(s) could not be proven contained
F-04	The harness suspected itself during this run
F-05	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: 78.7s.

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 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-20260907-30aad916
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 08:50:58	Evaluation run started
2026-09-07 08:52:17	Evaluation run finished
2026-09-07 08:52:17	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	0 / 1	1	0	1 / 1	78.5s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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: structured. **Rubric:** 1.0.

Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.

Submission 3647f6fe-2699-4011-8b84-36cca9fd7275

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/opencode-muse-spark	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/opencode-muse-spark	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

unverified_isolation × 1 (cli/opencode-muse-spark)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

No description recorded for this class.

F-02 · derived from the recorded run data

F-03 ATTENTION

1 rollout(s) could not be proven contained

Severity: ATTENTION | Evidence: Appendix A

No evidence establishes that the taker stayed inside the exam channel, and none establishes that it left. These rows are excluded from capability arithmetic and are neither passes nor model failures.

F-03 · derived from the recorded run data

F-04 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R4 containment unproven in 1/1 rollouts (lanes ['cli/opencode-muse-spark']) — no evidence the taker was confined to the exam channel. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-04 · derived from the recorded run data

F-05 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-05 · 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 ≥ 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-20260907T085058Z

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

9bcf06ca2f1d923f9b41bfb101fa96c67fa2cf601b2efd7a02d92c2e7279d19f

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

SOURCE RECORD FILE

82f4d2ce931a5db329a58bcb7fe062d349965a7ba90a3f63ae885c38cdcede7c

sha256 of fr-20260907-30aad916.json as written beside the standalone report.

SOURCE STANDALONE REPORT

800bfc37c101726a6dc3306dd91236d325f78f05536e0940dd4c4705fbd265c3

sha256 of fr-20260907-30aad916.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

252d43ddc7be671052890d48e3e8eb01ff00eb6b56c088ca342acab5c2fb0a68

sha256 of fr-20260907-30aad916.pdf.

RUN EVIDENCE BUNDLE

30aad916de820664fc88395c9979dbee95c19a90faf4a0aa03a4fe58032d85df

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-20260907-f20a63a7`. 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
3647f6fe	cli/opencode-muse-spark	0	submitted_passed	yes	Graded	fail	1	35 481	78.5s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
unverified_isolation	1	No description recorded for this class.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
30aad916de820664fc88395c9979dbee95c19a90faf4a0aa03a4fe58032d85df	live-20260907T085058Z	2026-09-07T08:50:58.543133+00:00	2026-09-07T08:52:17.292170+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	c2824b25572b9dce8a41a30b685a0eb1c65a2b036f8f1394d6cffbdc983ad7d5
rollouts/3647f6fe-2699-4011-8b84-36cca9fd7275.json	516eddf3291e8b5df8843391d2f2f4c7a65d968c4db224753a3b96907cb4d99d
spec	78c8729ef3f39ef7561f8c5cde9aabee41a18db6c3006a47ee1325fb88aca395
summary	9d1d31ff02f925ef5c7890b9ebb3e1c31e448865919fe23c3c32db675275ce45

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.

3647f6fe · cli:cli/opencode-muse-spark · 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	160
7	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.

14 Attempt record 2 of 3 · structured-data-1

`vvdex.annotation.structured-data-1 · record fr-20260907-34fc1989 · record digest 6e75dd966f50b3e2… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	1 of 1 graded rollouts	1 of 1 graded rollouts that called submit	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/opencode-muse-spark 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/opencode-muse-spark 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: 28.3s.

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 14520s 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-20260907-34fc1989
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 09:24:01	Evaluation run started
2026-09-07 09:24:30	Evaluation run finished
2026-09-07 09:24:30	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	1/1	0	0	1 / 1	28.0s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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 8fd52a1b-32f2-4d60-9162-f669da212743

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/opencode-muse-spark	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/opencode-muse-spark	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/opencode-muse-spark 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 ≥ 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-20260907T092401Z

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

6e75dd966f50b3e231b7ec7219e394760a634bc3a2a9b6ea6e80c60d39e1e737

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

SOURCE RECORD FILE

5d61670b128814b247452c276d515f2e3f2aaeb2099073a8370e3b1131bb2fb3

sha256 of fr-20260907-34fc1989.json as written beside the standalone report.

SOURCE STANDALONE REPORT

747315ede06f2ec129442847ec0e8835dc276fed7234c754f66812a7a30d9a2a

sha256 of fr-20260907-34fc1989.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

466a57c78bfcce480da3f698c2ed5c347dadafaab69e435a5848f1b40ea54b4a

sha256 of fr-20260907-34fc1989.pdf.

RUN EVIDENCE BUNDLE

34fc198967d140ad24b8f311bdb537f0243557498a91c89463a01c9e0ba9512b

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-20260907-30aad916`. 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
8fd52a1b	cli/opencode-muse-spark	0	submitted_passed	yes	Passed	pass	1	35 293	28.0s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
34fc198967d140ad24b8f311bdb537f0243557498a91c89463a01c9e0ba9512b	live-20260907T092401Z	2026-09-07T09:24:01.754994+00:00	2026-09-07T09:24:30.020451+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	7954878227170388db214d42762ae8d669f9771bfebc70109619e077822e54ba
rollouts/8fd52a1b-32f2-4d60-9162-f669da212743.json	3374b1009166f71ef4d6f9a7a244074ae424bed125f6054477284aab74732221
spec	d2987d367efabfb684862374a2b2eebeabf749277b40e00510703b22ddc617f2
summary	3416c45081fdf68ea1bbdae7ebb45e9101c5a1716af687bd30ff1cb2a946ca97

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.

8fd52a1b · cli:cli/opencode-muse-spark · 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	160
7	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.

15 Attempt record 3 of 3 · structured-data-1

`vvdex.annotation.structured-data-1 · record fr-20260907-f20a63a7 · record digest bcff1c08997a0869… · harness SUSPECT · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 0 graded rollouts	0 of 0 graded rollouts that called submit	1 of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam `vvdex.annotation.structured-data-1`. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 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 `model_api_failure` — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	model_api_failure × 1 (gemini-3.6-flash)
F-03	The harness suspected itself during this run
F-04	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: 528ms.

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 2280s 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-20260907-f20a63a7
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 08:45:29	Evaluation run started
2026-09-07 08:45:29	Evaluation run finished
2026-09-07 08:45:29	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.6-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	2280s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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: 0. Not graded — lane errors: 1. 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
gemini-3.6-flash `gemini-3.6-flash`	1	0	not graded	0	1	–	–	nothing

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
gemini-3.6-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; 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	GEMINI:GEMINI-3.6-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.6-flash	nothing	nothing	0 / 0	1

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.

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
gemini:gemini-3.6-flash	1	0 / 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
gemini:gemini-3.6-flash	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.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.6-flash	0	n/a	n/a	n/a — telemetry unavailable	0 / 0	0.0%

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 LANE / INFRA

model_api_failure × 1 (gemini-3.6-flash)

Severity: LANE / INFRA | Evidence: Appendix A · Appendix C

The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.

F-02 · derived from the recorded run data

F-03 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-03 · derived from the recorded run data

F-04 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-04 · 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

The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES

gemini-3.6-flash

WHAT WE WOULD LOOK AT FIRST

The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R–02 **For anyone reading the percentages**

Commission a run at n ≥ 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-20260907T084529Z

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

bcff1c08997a086933b905833db2746bf7460572b7642ccdc9b565917c352e34

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

SOURCE RECORD FILE

961110353da4d2fd2b4e306ac94f7fc36eca64e444fe6448600c4fda99c17c06

sha256 of fr-20260907-f20a63a7.json as written beside the standalone report.

SOURCE STANDALONE REPORT

9b281011de91807180bfb3580d79e11a9a8db6c7fb6663ab2a7090143c1d3a92

sha256 of fr-20260907-f20a63a7.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

530753b54e8d34bb0d0a3a02549ccce3b5bd3be5fa9c5e3ff4e180edc69d0f0b

sha256 of fr-20260907-f20a63a7.pdf.

RUN EVIDENCE BUNDLE

f20a63a712457a788717a03b8c86ae07f0ca51c7cc8b2ed7e4873685740d1c72

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
55e818c9	gemini-3.6-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	248ms

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
f20a63a712457a788717a03b8c86ae07f0ca51c7cc8b2ed7e4873685740d1c72	live-20260907T084529Z	2026-09-07T08:45:29.458015+00:00	2026-09-07T08:45:29.986350+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	a9fb1539ede1e85fdf84ca1ae3b06a4ad54986d268037171fad6c4d504ea4310
rollouts/55e818c9-c070-420b-b1f1-1e295c556a4d.json	042e2f655db2e8702be727f20e31e8e0417d15ec95be87ef8c8e270b5decf329
spec	97ca8ec04b58c4a6388026bfbcf52d341d8aac1439b0d6e6e15cede0ae73f3c7
summary	3e9f7b54872e5a8016fce4b1f3781c3c97c3c3bfdc7e40c1da2050b2ad59974f

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.

55e818c9 · gemini:gemini-3.6-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
No tool call was recorded for this rollout.				

Submission diff withheld by the public-disclosure policy. 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 1 of 3 · text-labeling-1

`vvdex.annotation.text-labeling-1 · record fr-20260907-020d48a8 · record digest 7d7ffa735d76c022… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	1 of 1 graded rollouts	1 of 1 graded rollouts that called submit	0 of 1 attempts

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/opencode-muse-spark 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/opencode-muse-spark 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 14520s 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-20260907-020d48a8
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 09:23:14	Evaluation run started
2026-09-07 09:24:01	Evaluation run finished
2026-09-07 09:24:01	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	1/1	0	0	1 / 1	47.3s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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 262771ab-b919-41f5-825e-dc988b90d10e

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/opencode-muse-spark	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/opencode-muse-spark	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/opencode-muse-spark 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 ≥ 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-20260907T092314Z

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

7d7ffa735d76c02249d6f8a115660a1873f1b2c39cafe2af4dacd4afddd74db8

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

SOURCE RECORD FILE

676fc302738c93eea9bbf83e011dc0cb77b552415386a0e7c756220c4fa91735

sha256 of fr-20260907-020d48a8.json as written beside the standalone report.

SOURCE STANDALONE REPORT

329df4b34888116aaf627776ec8634c8e688d670dd5d219024dd3465527c8116

sha256 of fr-20260907-020d48a8.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

f442801137db76380997dde3ae6aff755a147e0af13a52611d884f36c068e12c

sha256 of fr-20260907-020d48a8.pdf.

RUN EVIDENCE BUNDLE

020d48a88ce3b37ca360343b8da6c083434eec1edc5c4be1b60e89965a37ebaf

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-20260907-f200668c`. 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
262771ab	cli/opencode-muse-spark	0	submitted_passed	yes	Passed	pass	1	34 221	47.3s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
020d48a88ce3b37ca360343b8da6c083434eec1edc5c4be1b60e89965a37ebaf	live-20260907T092314Z	2026-09-07T09:23:14.100605+00:00	2026-09-07T09:24:01.637200+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	17d8a583122975b8be0d88fddf161a750615d3d28cd5fa89dbd74963719b2072
rollouts/262771ab-b919-41f5-825e-dc988b90d10e.json	6fb72b9e0bc184a2ba39c3ad9348c76a72942c35369faa648716be84e19fdaf9
spec	ae7d630f3f37b90571d50026baa8cab9fe7f1cb20d285772ce919545ff3628b9
summary	022b2d274557a149b24cb42b3ce382f545604ff3d85950b62e7d601184c4df20

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.

262771ab · cli:cli/opencode-muse-spark · 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 (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.

17 Attempt record 2 of 3 · text-labeling-1

`vvdex.annotation.text-labeling-1 · record fr-20260907-55873dc4 · record digest 668ec8bd2ec34cac… · harness SUSPECT · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 0 graded rollouts	0 of 0 graded rollouts that called submit	1 of 1 attempts

Purpose

This report records 1 interrupted attempt(s) against the certified exam `vvdex.annotation.text-labeling-1`. No submission reached the grader. The frozen exam defines the workspace, tool contract and limits; this record documents where execution stopped, not a graded model result.

Outcome

0 of 0

graded rollouts passed the independent hidden tests and were certified — 1 of 1 attempts not graded (lane errors)

0 of 0 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 `model_api_failure` — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	model_api_failure × 1 (gemini-3.6-flash)
F-03	The harness suspected itself during this run
F-04	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: 507ms.

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 2280s 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-20260907-55873dc4
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 08:45:28	Evaluation run started
2026-09-07 08:45:29	Evaluation run finished
2026-09-07 08:45:29	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
gemini-3.6-flash	gemini	VVDex-configured	operator API key; billing tier and monetary charge not recorded

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	2280s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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: 0. Not graded — lane errors: 1. 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
gemini-3.6-flash `gemini-3.6-flash`	1	0	not graded	0	1	–	–	nothing

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
gemini-3.6-flash	1	0	1	0	not recorded	[0%, 100%]	0

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = ; 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	GEMINI:GEMINI-3.6-FLASH
Inspect	0 / 0
Annotate	0 / 0
Validate	0 / 0
Submit	0 / 0
Graded	0 / 0
Passed	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
gemini:gemini-3.6-flash	nothing	nothing	0 / 0	1

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.

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
gemini:gemini-3.6-flash	1	0 / 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
gemini:gemini-3.6-flash	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.

Efficiency & telemetry

This run has no per-token price attached — a local model, a free quota, or a customer endpoint we are not billed for — so spend is reported in tokens. Tokens are what was actually measured; a dollar figure here would be invented. Lanes run through a flat-rate local CLI report no token telemetry; their cells say n/a — an absent measurement, not zero. Rollout counts here are evaluable rollouts.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
gemini:gemini-3.6-flash	0	n/a	n/a	n/a — telemetry unavailable	0 / 0	0.0%

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 LANE / INFRA

model_api_failure × 1 (gemini-3.6-flash)

Severity: LANE / INFRA | Evidence: Appendix A · Appendix C

The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.

F-02 · derived from the recorded run data

F-03 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R1 api-failure share 1/1 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-03 · derived from the recorded run data

F-04 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-04 · 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

The model endpoint failed to answer. Not a verdict about the model's ability.

AFFECTED LANES

gemini-3.6-flash

WHAT WE WOULD LOOK AT FIRST

The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

R–02 **For anyone reading the percentages**

Commission a run at n ≥ 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-20260907T084528Z

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

668ec8bd2ec34cacaff740b1495750d90b0052a493bb9f1b15a2b10341870c78

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

SOURCE RECORD FILE

9c02cfd55c1b31b1b67b75160488a99f412688ef67eb3ae5766fe27f0c422e73

sha256 of fr-20260907-55873dc4.json as written beside the standalone report.

SOURCE STANDALONE REPORT

e62e4c61eb1b0b6da141881f44763bd788b06f21304b29c968bd8a9c7a872efe

sha256 of fr-20260907-55873dc4.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

6b7b690430cd73add44e49935d4399356aabab873c0b25df340b0037b67381f0

sha256 of fr-20260907-55873dc4.pdf.

RUN EVIDENCE BUNDLE

55873dc4697368e0e41ebcfdff05a429e5c11dc069553812a351a7f797126aa1

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

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
39a85b1f	gemini-3.6-flash	0	model_api_failure	no	Not graded — lane error	not graded	0	n/a	273ms

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
55873dc4697368e0e41ebcfdff05a429e5c11dc069553812a351a7f797126aa1	live-20260907T084528Z	2026-09-07T08:45:28.861487+00:00	2026-09-07T08:45:29.369347+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	37d7b54116b9ef91928b20e2c356c308eaaef0fa293d7f12131f99fced29b435
rollouts/39a85b1f-4236-4f16-9a3c-1adfbd2edbdf.json	29939ccce2968de54fa3b346403560c7b15c3bd1dd726e15ef1a1f53ac5fd39d
spec	4d6c9dfb892c9c4d0f80a9702ffb65bd3c9be7ed04a7fe938e7443b33e24a301
summary	b6c529fb36e235cb5bf2a28f629f872074040f04056219261e2dd3c1c5ceb62d

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.

39a85b1f · gemini:gemini-3.6-flash · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
No tool call was recorded for this rollout.				

Submission diff withheld by the public-disclosure policy. 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 3 of 3 · text-labeling-1

`vvdex.annotation.text-labeling-1 · record fr-20260907-f200668c · record digest a8f1f033b396a079… · harness SUSPECT · {"breached": 0, "unverified": 1}`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	0 of 1 graded rollouts	1 of 1 graded rollouts that called submit	0 of 1 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vv`dex.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

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 `unverified_isolation` — .

Key findings

F-01	No evaluated lane passed the hidden tests in this run
F-02	unverified_isolation × 1 (cli/opencode-muse-spark)
F-03	1 rollout(s) could not be proven contained
F-04	The harness suspected itself during this run
F-05	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: 164.6s.

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 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-20260907-f200668c
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 08:48:13	Evaluation run started
2026-09-07 08:50:58	Evaluation run finished
2026-09-07 08:50:58	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Muse Spark 1.3 via OpenCode CLI (free contributor tier, 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	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	24	14520s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	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
Muse Spark 1.3 via OpenCode CLI (free contributor tier, image input) `cli/opencode-muse-spark`	1	1	0 / 1	1	0	1 / 1	164.3s	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/opencode-muse-spark	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/OPENCODE-MUSE-SPARK
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/opencode-muse-spark	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: text. **Rubric:** 1.0.

An invented message is reviewed for classification and entity spans with consistent offsets.

Submission e6f66def-d4b6-47fe-b9af-0fb217f59a1b

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/opencode-muse-spark	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/opencode-muse-spark	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

unverified_isolation × 1 (cli/opencode-muse-spark)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

No description recorded for this class.

F-02 · derived from the recorded run data

F-03 ATTENTION

1 rollout(s) could not be proven contained

Severity: ATTENTION | Evidence: Appendix A

No evidence establishes that the taker stayed inside the exam channel, and none establishes that it left. These rows are excluded from capability arithmetic and are neither passes nor model failures.

F-03 · derived from the recorded run data

F-04 NOTE

The harness suspected itself during this run

Severity: NOTE | Evidence: Appendix A · Appendix C

Self-suspicion rules fired: R4 containment unproven in 1/1 rollouts (lanes ['cli/opencode-muse-spark']) — no evidence the taker was confined to the exam channel. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.

F-04 · derived from the recorded run data

F-05 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-05 · 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 ≥ 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-20260907T084813Z

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

a8f1f033b396a079639a3b4c4147e651b099ff45a7ac4a2a8f0d60d253139572

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

SOURCE RECORD FILE

66234de9cc7e376f7f913bfb1636e2e9142a231ece91129f29e8ae4484ae62ea

sha256 of fr-20260907-f200668c.json as written beside the standalone report.

SOURCE STANDALONE REPORT

74ef111d7a569424d304199d228b9d712a117568288e89a240c5d053079ff646

sha256 of fr-20260907-f200668c.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

075f6b5f7fbd4a24d05b86ee1c35dc1ebb18d839f2b9187d9bfe0a5d81f61220

sha256 of fr-20260907-f200668c.pdf.

RUN EVIDENCE BUNDLE

f200668cb308b4fcd627c1f4211a525b71f77532389dd47288397f2eb10bf20b

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-20260907-55873dc4`. 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
e6f66def	cli/opencode-muse-spark	0	submitted_passed	yes	Graded	fail	1	30 640	164.3s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
unverified_isolation	1	No description recorded for this class.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
f200668cb308b4fcd627c1f4211a525b71f77532389dd47288397f2eb10bf20b	live-20260907T084813Z	2026-09-07T08:48:13.866977+00:00	2026-09-07T08:50:58.428453+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	71ab7b7a72404960edfb6277be5c743032f573a0fc65545495875914e7067f70
rollouts/e6f66def-d4b6-47fe-b9af-0fb217f59a1b.json	6abbaa80b25d06b70557970e7127ebb1435ec82a350416a79b1aa72f772377e3
spec	9c72c384afa2df97c220fba3396d9e371e7d9c292ad12bffa66a92e7d2edf74a
summary	c6ebb0d16a31847630bf078deaf6e994a140c29a50a020ce7e753dfcec741f79

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.

e6f66def · cli:cli/opencode-muse-spark · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
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 (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.

19 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.audio-events-1 → fr-20260907-67c07cce · recordDigest 1f595aa9846136ff6fcd075c0a6f491a969633f119bb0c63505f70a92e0e6a02
- vvdex.annotation.image-object-1 → fr-20260907-54b8ff2d · recordDigest 020983794ee70906ee3c9162e3326cb5575cbf75cd4963d811d1378a966c732c
- vvdex.annotation.image-object-1 → fr-20260907-9bee4b48 · recordDigest b08ceac9eef6a1d09760e410f2b09a12c6de4c221c95cdc8aef20a6fc5ef776f
- vvdex.annotation.image-object-1 → fr-20260907-e2134530 · recordDigest 80e72cacb9087660e9efeab8e1ab8c7635a7463c8d2b7be1d7900839ea280dd8
- vvdex.annotation.structured-data-1 → fr-20260907-30aad916 · recordDigest 9bcf06ca2f1d923f9b41bfb101fa96c67fa2cf601b2efd7a02d92c2e7279d19f
- vvdex.annotation.structured-data-1 → fr-20260907-34fc1989 · recordDigest 6e75dd966f50b3e231b7ec7219e394760a634bc3a2a9b6ea6e80c60d39e1e737
- vvdex.annotation.structured-data-1 → fr-20260907-f20a63a7 · recordDigest bcff1c08997a086933b905833db2746bf7460572b7642ccdc9b565917c352e34
- vvdex.annotation.text-labeling-1 → fr-20260907-020d48a8 · recordDigest 7d7ffa735d76c02249d6f8a115660a1873f1b2c39cafe2af4dacd4afddd74db8
- vvdex.annotation.text-labeling-1 → fr-20260907-55873dc4 · recordDigest 668ec8bd2ec34cacaff740b1495750d90b0052a493bb9f1b15a2b10341870c78
- vvdex.annotation.text-labeling-1 → fr-20260907-f200668c · recordDigest a8f1f033b396a079639a3b4c4147e651b099ff45a7ac4a2a8f0d60d253139572

20 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.