

Muse Spark 1.3 via OpenCode CLI — text, structured and image, ten attempts each

3 fixed tasks · 3 sealed attempt records · 30 graded attempts · 20 certified passes

CURRENT

CAMPAIGN REPORT

2026-09-07

Evaluation scope:

cli/opencode-muse-spark

Prepared by:

VVDex Forge — regenerated from verified sealed records only

fc-e4d7e6adab0f · 3 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	4
03.	Environment families	5
04.	Campaign aggregate across these three exams	6
05.	Harness status	7
06.	Cells excluded from capability	8
07.	What the outcomes mean	9
08.	What this report withholds, and why	10
09.	Attempt record 1 of 1 · image-object-1	11
10.	Attempt record 1 of 1 · structured-data-1	33
11.	Attempt record 1 of 1 · text-labeling-1	55
12.	Evidence	77
13.	Attestation	78

01 Executive summary

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

FIXED TASKS 3	SEALED RECORDS 3	MODEL LANES 1	ATTEMPTS 30
GRADED ATTEMPTS 30	CERTIFIED PASSES 20	LANE ERRORS (NOT GRADED) 0	WITHHELD / INVALID 0

Purpose

This report presents a controlled evaluation of 1 model lane(s) against 3 fixed certified annotation tasks, one sealed attempt record on each task. For each task, every lane received that task's frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw; and every rollout carries three separate outcomes — grader, integrity, certified — of which only the certified outcome is counted. Each sealed attempt record is rendered below as a chapter derived from the same record as its standalone report: the chapter's bytes are authenticated by this campaign document, the standalone report by its own digest.

Annotation campaign scope. This campaign evaluates three fixed annotation tasks with one sealed attempt record on each task (3 sealed attempts in total). The sealed tasks cover image, structured, text inputs. Audio is N/A because no audio task appears in this campaign's sealed records. CLI temperature and stochastic-seed controls are not established by these records; attempt indices are bookkeeping. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

20

certified pass(es) across 30 graded attempts (30 attempts, 0 lane error(s) not graded) in 3 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-e4d7e6adab0f · generated from 3 verified sealed record(s)

Key findings across the campaign

- image-object-1 · cli/opencode-muse-spark · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · cli/opencode-muse-spark · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/opencode-muse-spark · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION · IMAGE-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	0/10	10/10	10/10
-------------------------	------	-------	-------

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

03 Environment families

This campaign covers three 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.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 three exams

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
cli/opencode-muse-spark	20	30	0	0	0

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

05 Harness status

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

06 Cells excluded from capability

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

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

07 What the outcomes mean

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

08 What this report withholds, and why

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

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

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

vvdex.annotation.image-object-1 · record fr-20260907-d8051665 · record digest 533b6d1f4d350c54... · harness clean · containment clean

Executive summary

MODELS TESTED 1	CERTIFIED PASSES 0 of 10 graded rollouts	SUBMITTED 10 of 10 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 10 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 10

graded rollouts passed the independent hidden tests and were certified

0 of 10 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. 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 × 10 (cli/opencode-muse-spark)

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: 2025.8s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 14520s of wall clock, at 10 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

The instrument. Exam `vvdex.annotation.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	10 (10 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfde31a0ce8ae2d7295

EXAM FINGERPRINT

422753713aeb9f41abcc4135...

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

Evaluation configuration

REPORT	fr-20260907-d8051665
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 10:09:14	Evaluation run started
2026-09-07 10:43:00	Evaluation run finished
2026-09-07 10:43:00	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 docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 10. Graded (evaluable) rollouts: 10. 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	10	10	0 / 10	10	0	10 / 10	202.3s	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	PASS@3	PASS@5	PASS@10
cli/opencode-muse-spark	10	10	0	0	0.0%	[0%, 28%]	0–9	0.0%	0.0%	0.0%	0.0%

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

Model comparison

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	10 / 10
Annotate	10 / 10
Validate	10 / 10
Submit	10 / 10
Graded	10 / 10
Passed	0 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/opencode-muse-spark	Graded	Graded	0 / 10	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 9db02f79-274f-45a0-a7cf-01688f594b74

Score	Items	Valid / invalid	Types
0.7698	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.5521
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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
------------------	--------------

Submission 98dc8366-ab16-4e8a-8248-0a8744219763

Score	Items	Valid / invalid	Types
0.7415	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.4955
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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

Submission cd13432b-0846-4185-838b-f137051967d2

Score	Items	Valid / invalid	Types
0.7415	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.4955
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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

Submission f9fb75d2-0eec-4ba0-81ab-695ee86a4636

Score	Items	Valid / invalid	Types
0.7389	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.4955
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.0354	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

Submission 1ee5c074-df9c-4794-afbf-3fedbc6f95a6

Score	Items	Valid / invalid	Types
0.7107	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.4391
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.0354	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

Submission 49ed0958-99cc-4547-aeb4-c318448239d5

Score	Items	Valid / invalid	Types
0.7529	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.5239
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.0361	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

Submission 5799a6ce-4c73-4f24-87bc-28d78bbf8a34

Score	Items	Valid / invalid	Types
0.775	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	1	boxIoU	0.5625
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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

Submission a415324f-6938-49f1-ad2c-52d5ae792312

Score	Items	Valid / invalid	Types
0.7415	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.4955
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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

Submission f167b53b-0192-475d-bd23-e232731d9262

Score	Items	Valid / invalid	Types
0.775	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	1	boxIoU	0.5625
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.025	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

Submission b9192f49-2bc8-4a99-8d26-fd0abba684f6

Score	Items	Valid / invalid	Types
0.7475	4	4 / 0	box, classification, keypoint

Measured metrics

annotationCount	4	boxF1	0.5	boxIoU	0.5091
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
keypointF1	1	keypointNormalizedError	0.0283	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	0.6	–	–	
boxRecall	recorded	N/A	0.6	–	–	
boxF1	recorded	N/A	0.6	–	–	
boxIoU	recorded	N/A	0.5131	–	–	
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)

0.7494

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	10	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	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 10 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 × 10 (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

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.

Verification

RUN ID

live-20260907T100914Z

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

533b6d1f4d350c544088d42f0a3f3f3943d6e17c1095ac620bfe52406a8aeb5b

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

SOURCE RECORD FILE

2ad2999b67a4ed31e328f41f43ba2dd3b1f36c5576b8a42c7a018d818b7fd011

sha256 of `fr-20260907-d8051665.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

1ac9fbd95b4e001d544aaff8a2a3c00e36f5fb94756658ec45aad65d228eb959

sha256 of `fr-20260907-d8051665.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

407cf7248d9096c04724132315ee9e58d5e9b7cf429e18876852531bd3ec55a8

sha256 of `fr-20260907-d8051665.pdf`.

RUN EVIDENCE BUNDLE

d805166565c67c2442e97abf44b88a0c5a96cf622f1cba076cd46a1b1ec9cb3b

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

CERTIFIED EVIDENCE SEAL

cf945b4838a266664300b32fdd5fed0c4e781ab95eaaebfd e31a0ce8ae2d7295

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

Measurement history

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
9db02f79	cli/opencode-muse-spark	0	submitted_failed	yes	Graded	fail	1	35 694	185.6s
98dc8366	cli/opencode-muse-spark	1	submitted_failed	yes	Graded	fail	1	39 100	178.7s
cd13432b	cli/opencode-muse-spark	2	submitted_failed	yes	Graded	fail	1	36 564	109.0s
f9fb75d2	cli/opencode-muse-spark	3	submitted_failed	yes	Graded	fail	1	36 157	209.3s
1ee5c074	cli/opencode-muse-spark	4	submitted_failed	yes	Graded	fail	1	27 957	254.8s
49ed0958	cli/opencode-muse-spark	5	submitted_failed	yes	Graded	fail	1	36 518	263.6s
5799a6ce	cli/opencode-muse-spark	6	submitted_failed	yes	Graded	fail	1	39 149	252.1s
a415324f	cli/opencode-muse-spark	7	submitted_failed	yes	Graded	fail	1	38 107	220.6s
f167b53b	cli/opencode-muse-spark	8	submitted_failed	yes	Graded	fail	1	38 095	213.3s
b9192f49	cli/opencode-muse-spark	9	submitted_failed	yes	Graded	fail	1	39 067	136.4s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	10	Submitted annotations rejected by the independent grader.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

d805166565c67c2442e97abf44b88a0c5a96cf622f1cba076cd46a1b1ec9cb3b

RUN ID

live-20260907T100914Z

STARTED

2026-09-07T10:09:14.524595+00:00

FINISHED

2026-09-07T10:43:00.368891+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	6110297e106b059c2ca05d995532bcc440bf2126ad9dc8f9982493b1906970ee
rollouts/1ee5c074-df9c-4794-afbf-3fedbc6f95a6.json	e1fc0ccd4713752b55263244180acbf7f2984110df3bf41ce0c665621314ad87
rollouts/49ed0958-99cc-4547-aeb4-c318448239d5.json	72fca7de8404fadb3bff739418f0fd51fff4c1df6730260eb76dd15f872a52a1
rollouts/5799a6ce-4c73-4f24-87bc-28d78bbf8a34.json	2da3bde341d6d99dc96def128eb26e8cb75b053d1e2051d45ab98a662588db09
rollouts/98dc8366-ab16-4e8a-8248-0a8744219763.json	dd3fe7f45179a91d78c967e41da515086cbca6509f9d18f7692a27186bf9241a
rollouts/9db02f79-274f-45a0-a7cf-01688f594b74.json	ca4d241eb5cd7eaeef27bb50c3cb4a2907e86e452e517b5dcd7eb60fbbcd56ff0
rollouts/a415324f-6938-49f1-ad2c-52d5ae792312.json	69f017f00dd9254dc37b515949b62346406d04a9d9f56201b2b247bbcb00c4b9
rollouts/b9192f49-2bc8-4a99-8d26-fd0abba684f6.json	451a49ea697a01b39623ca71ccf44ef5da11a04066e5f52e47271a2f2ffc6012
rollouts/cd13432b-0846-4185-838b-f137051967d2.json	bfc2594d60dba0a37bd4a881a37ba9bbb6b009bebecd37ebbfeba04b44ecec9
rollouts/f167b53b-0192-475d-bd23-e232731d9262.json	e5174f4c1d20e020f4ab503e1c01253c2aa828f8c05f4d2fb9f0a1cda9020d08
rollouts/f9fb75d2-0eec-4ba0-81ab-695ee86a4636.json	0c7dba33308b8c6127a0cdf6d3836723880bc178ed5df342e5ceb4c1f12c422a
spec	83ed87130dfcc1b2449c947afa273e623554aa820b3e6d07e539ef6bb64cfe2c
summary	7e01bf349aefcdcc2e70fb7b43b530ea02bfe8d14016f04e6a54a8d7a0d938ba

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.

9db02f79 · 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.

98dc8366 · cli:cli/opencode-muse-spark · seed 1 · 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.

cd13432b · cli:cli/opencode-muse-spark · seed 2 · 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.

f9fb75d2 · cli:cli/opencode-muse-spark · seed 3 · 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.

1ee5c074 · cli:cli/opencode-muse-spark · seed 4 · 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.

49ed0958 · cli:cli/opencode-muse-spark · seed 5 · 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.

5799a6ce · cli:cli/opencode-muse-spark · seed 6 · 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.

a415324f · cli:cli/opencode-muse-spark · seed 7 · 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

STEP	TOOL	PATH	RESULT	CHARS
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

f167b53b · cli:cli/opencode-muse-spark · seed 8 · 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.

b9192f49 · cli:cli/opencode-muse-spark · seed 9 · 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.

10 Attempt record 1 of 1 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260907-4b1e97bc · record digest 72451cce4189c11a... · harness clean · c
ontainment clean

Executive summary

MODELS TESTED 1	CERTIFIED PASSES 10 of 10 graded rollouts	SUBMITTED 10 of 10 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 10 attempts
---------------------------------------	--	--	--

Purpose

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

Outcome

10 of 10

graded rollouts passed the independent hidden tests and were certified

10 of 10 graded rollouts passed. cli/opencode-muse-spark completed the applicable
annotation workflow and were accepted by the independent annotation grader.

Key findings

F-01 cli/opencode-muse-spark passed the independent hidden tests

Each finding is stated in full, with evidence, in the Findings section; recommendations for
acting on them follow it. Elapsed wall clock for the whole engagement: 1173.3s.

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 10 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.structured-data-1	0.1.0	annotation	10 (10 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

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

Evaluation configuration

REPORT	fr-20260907-4b1e97bc
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 09:49:40	Evaluation run started
2026-09-07 10:09:14	Evaluation run finished
2026-09-07 10:09:14	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 docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 10. Graded (evaluable) rollouts: 10. 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	10	10	10 / 10	0	0	10 / 10	117.1s	Passed

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
cli/opencode-muse-spark	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%

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

Model comparison

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	10 / 10
Annotate	10 / 10
Validate	10 / 10
Submit	10 / 10
Graded	10 / 10
Passed	10 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/opencode-muse-spark	Passed	Passed	10 / 10	0

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

Annotation evidence

Modality: structured. Rubric: 1.0.

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

Submission a7f03747-8bea-47e8-a8c3-24076a0b210b

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

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

Submission aee28375-f4c5-40a7-87f5-88698db65dfa

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

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

Submission eaeafc14-dc3f-4460-859b-582c2802f6d6

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

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

Submission b572d19c-adb6-4993-b180-093589c9ad3b

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

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

Submission 0888f237-fbf9-4620-ae97-666ed895c6bc

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics					
annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
None recorded					
Disagreement					
Agreement review				Not measured	

Submission cccf2d79-7048-4295-b260-7b97bca09577

Score	Items	Valid / invalid		Types	
1	4	4 / 0		pair, row	
Measured metrics					
annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
None recorded					
Disagreement					
Agreement review				Not measured	

Submission 4f5caecf-47f4-4fc4-973e-bc63b1acf17e

Score	Items	Valid / invalid	Types		
1	4	4 / 0	pair, row		
Measured metrics					
annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1

pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

Agreement reviewNot measured

Submission a56c86e8-f990-410e-afdb-a86a9e3e373e

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

Agreement reviewNot measured

Submission ab0e4a61-88ef-404f-ae62-29cc586aa20d

Score	Items	Valid / invalid	Types
1	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes	
None recorded	
Disagreement	
Agreement review	Not measured

Submission 7e24d987-7258-4da5-837e-f59f2b3d2e41

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	10	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	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 10 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

10 of 10 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

Recommendations

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

Verification

RUN ID

live-20260907T094940Z

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

72451cce4189c11abcdcbb77e3dfc4c37f13375d9be7ea006984415d15d5bbc

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

SOURCE RECORD FILE

3b00ed7ed175ff830ae51937ab36d8b1500267fbc406f00ff2205277a8d1f395

sha256 of fr-20260907-4b1e97bc.json as written beside the standalone report.

SOURCE STANDALONE REPORT

dc76ff1a92d010987d1c92387caafd7a3cc3b6f41f65654287d58bd2482f121a

sha256 of fr-20260907-4b1e97bc.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

a0f6d7754a70a18085cbc32e939ca23bf558dd5090230f013de1b7c3073bbeb5

sha256 of fr-20260907-4b1e97bc.pdf.

RUN EVIDENCE BUNDLE

4b1e97bc61f6313a302a90639255440b7d458a4fdbaff542ffa1d3a7ca9df4e7

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
a7f03747	cli/opencode-muse-spark	0	submitted_passed	yes	Passed	pass	1	33 437	107.8s
aee28375	cli/opencode-muse-spark	1	submitted_passed	yes	Passed	pass	1	33 269	71.5s
eaeafc14	cli/opencode-muse-spark	2	submitted_passed	yes	Passed	pass	1	32 929	116.1s
b572d19c	cli/opencode-muse-spark	3	submitted_passed	yes	Passed	pass	1	33 823	60.0s
0888f237	cli/opencode-muse-spark	4	submitted_passed	yes	Passed	pass	1	35 706	72.7s
cccf2d79	cli/opencode-muse-spark	5	submitted_passed	yes	Passed	pass	1	28 396	323.6s
4f5caecf	cli/opencode-muse-spark	6	submitted_passed	yes	Passed	pass	1	31 264	156.6s
a56c86e8	cli/opencode-muse-spark	7	submitted_passed	yes	Passed	pass	1	33 648	36.8s
ab0e4a61	cli/opencode-muse-spark	8	submitted_passed	yes	Passed	pass	1	35 934	131.0s
7e24d987	cli/opencode-muse-spark	9	submitted_passed	yes	Passed	pass	1	33 207	95.0s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

4b1e97bc61f6313a302a9063

9255440b7d458a4fdbaff542

ffa1d3a7ca9df4e7

RUN ID

live-

20260907T094940Z

STARTED

2026-09-

07T09:49:40.938019

+00:00

FINISHED

2026-09-

07T10:09:14.281105

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	a1a86644ac2b6865f46d30e40dd325f91082d6db734c407b5b9b8c98058f65b4
rollouts/0888f237-fbf9-4620-ae97-666ed895c6bc.json	5dafbfdc69a9ce0310c48e2fd02620e7b3f8aec3c6fe471d10ac82373a98b38e
rollouts/4f5caecf-47f4-4fc4-973e-bc63b1acf17e.json	d1f098842340a028e6abe11c99591b6628c18c8fcfe5db20cf0b224570d3d75c
rollouts/7e24d987-7258-4da5-837e-f59f2b3d2e41.json	35d5fd8c28e2dc1c2b0ab7ee86a50205d030db54cedb62de29012478395bc1c6
rollouts/a56c86e8-f990-410e-afdb-a86a9e3e373e.json	9750fcf56d84fbc87bec21270d8ec4dd4cbe38956db171b04a4b997c9639f318
rollouts/a7f03747-8bea-47e8-a8c3-24076a0b210b.json	9424d8ad0b1475e1ecaa6619871e762a41639469a71b2fdb0a23aeca143976c9
rollouts/ab0e4a61-88ef-404f-ae62-29cc586aa20d.json	0454a77ebc7dcdc155ecc3c05035b30c295b52ebc016391007421793931c89ad
rollouts/aee28375-f4c5-40a7-87f5-88698db65dfa.json	64cd99711468d46734d8af080bd74c16d7e28e06ed71aaeb3862bcb11bc77c00
rollouts/b572d19c-adb6-4993-b180-093589c9ad3b.json	21d40a52daba1481797dc150c0f3c8c8d416eec3b2cdbe669106fb446679eb56
rollouts/cccf2d79-7048-4295-b260-7b97bca09577.json	3154f85b297ce3072eb83433374a130930ce47fcad5540a460bc04443d430c64
rollouts/eaefc14-dc3f-4460-859b-582c2802f6d6.json	aa5a5cabdc9606d31706098623cd8227b327cf50fbd475c01be462a6ae59baf2
spec	838264ae864e1f4c7142f5d3e54362651021df23921a65ff03f4ce13b227097e
summary	935b1f35bdc4ab38d96994b46074db5a99f2fb1380ac98a174170e6f5120f923

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.

a7f03747 · 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.

aee28375 · cli:cli/opencode-muse-spark · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

eaeafc14 · cli:cli/opencode-muse-spark · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

b572d19c · cli:cli/opencode-muse-spark · seed 3 · 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.

0888f237 · cli:cli/opencode-muse-spark · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (38 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

cccf2d79 · cli:cli/opencode-muse-spark · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

4f5caecf · cli:cli/opencode-muse-spark · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

a56c86e8 · cli:cli/opencode-muse-spark · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74

STEP	TOOL	PATH	RESULT	CHARS
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

ab0e4a61 · cli:cli/opencode-muse-spark · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	submission.json	ok	74
4	read_file	assets/records.json	ok	295
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.

7e24d987 · cli:cli/opencode-muse-spark · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	list_files	.	ok	81
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.

11 Attempt record 1 of 1 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-c308bc27 · record digest 89f0a9f902db85a5... · harness clean · containment clean

Executive summary

MODELS TESTED 1	CERTIFIED PASSES 10 of 10 graded rollouts	SUBMITTED 10 of 10 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 10 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

10 of 10

graded rollouts passed the independent hidden tests and were certified

10 of 10 graded rollouts passed. cli/opencode-muse-spark completed the applicable annotation workflow and were accepted by the independent annotation grader.

Key findings

F-01 cli/opencode-muse-spark passed the independent hidden tests

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 1494.6s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 14520s of wall clock, at 10 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.text-labeling-1	0.1.0	annotation	10 (10 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

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

Evaluation configuration

REPORT	fr-20260907-c308bc27
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 09:24:46	Evaluation run started
2026-09-07 09:49:40	Evaluation run finished
2026-09-07 09:49:40	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 docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 10. Graded (evaluable) rollouts: 10. 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	10	10	10 / 10	0	0	10 / 10	149.2s	Passed

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
cli/opencode-muse-spark	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%

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

Model comparison

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	10 / 10
Annotate	10 / 10
Validate	10 / 10
Submit	10 / 10
Graded	10 / 10
Passed	10 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/opencode-muse-spark	Passed	Passed	10 / 10	0

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

Annotation evidence

Modality: text. Rubric: 1.0.

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

Submission 7c8b66f3-3ce4-410d-b1a8-6387f5c07369

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

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

Submission c3604498-c12c-42a1-bdf2-8ba54361ef3a

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

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

Submission 2d1074f7-1e2b-4b54-a7f5-75932f3edd34

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

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

Submission 18e71903-405f-4fdd-9e86-d26c7999d977

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

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

Submission a3129789-0993-492e-8df4-88f31b25cf5d

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

Agreement review

Not measured

Submission 028ac44d-71c0-4be5-a235-ad2e1697353d

Score	Items	Valid / invalid		Types	
1	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

Agreement review

Not measured

Submission c2ddef0b-20d0-4603-92f4-38e6af50c4ef

Score	Items	Valid / invalid		Types	
1	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3

mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

Agreement reviewNot measured

Submission 2cc0a656-e29c-4d64-8525-f42386d9b497

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes

None recorded

Disagreement

Agreement reviewNot measured

Submission 358e1aa0-da0a-4421-a6f5-67f4ecc59dd6

Score	Items	Valid / invalid	Types
1	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	0	spanExactF1	1	spanF1	1
spanOverlapIoU	1	spanPrecision	1	spanRecall	1

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

Validation error codes	
None recorded	
Disagreement	
Agreement review	Not measured

Submission 31f2d57e-85b9-47ed-8440-29acc6ea489b

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	10	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	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 10 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

10 of 10 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

Recommendations

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

Verification

RUN ID

live-20260907T092446Z

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

89f0a9f902db85a59371bfb233988ed854547dc96e9c1a2a8d6734bbf2943960

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

SOURCE RECORD FILE

00aeaade4baed5fcd763dc075956b53ceb0b4cb93f1a973ab8020f6b62d98a4e

sha256 of `fr-20260907-c308bc27.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

c2c99536959af8aac6f0efb0499ceb3dfcef9594d9972e3b47ed0fc957e01b13

sha256 of `fr-20260907-c308bc27.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

edab6da632ab08705069e013e327ae6f830d97076243b87b00c92987fce33f8d

sha256 of `fr-20260907-c308bc27.pdf`.

RUN EVIDENCE BUNDLE

c308bc278efb208a747ec20e09790e345daabac81683c3949351be442b97b2ed

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

CERTIFIED EVIDENCE SEAL

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282

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

Measurement history

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

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

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
7c8b66f3	cli/opencode-muse-spark	0	submitted_passed	yes	Passed	pass	1	31 607	192.7s
c3604498	cli/opencode-muse-spark	1	submitted_passed	yes	Passed	pass	1	26 880	207.2s
2d1074f7	cli/opencode-muse-spark	2	submitted_passed	yes	Passed	pass	1	34 246	136.5s
18e71903	cli/opencode-muse-spark	3	submitted_passed	yes	Passed	pass	1	33 776	72.5s
a3129789	cli/opencode-muse-spark	4	submitted_passed	yes	Passed	pass	1	28 538	231.5s
028ac44d	cli/opencode-muse-spark	5	submitted_passed	yes	Passed	pass	1	34 656	77.7s
c2ddef0b	cli/opencode-muse-spark	6	submitted_passed	yes	Passed	pass	1	34 431	145.5s
2cc0a656	cli/opencode-muse-spark	7	submitted_passed	yes	Passed	pass	1	34 369	161.4s
358e1aa0	cli/opencode-muse-spark	8	submitted_passed	yes	Passed	pass	1	34 419	123.4s
31f2d57e	cli/opencode-muse-spark	9	submitted_passed	yes	Passed	pass	1	33 724	144.1s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

c308bc278efb208a747ec20e09790e345daabac81683c3949351be442b97b2ed

RUN ID

live-20260907T092446Z

STARTED

2026-09-07T09:24:46.212632+00:00

FINISHED

2026-09-07T09:49:40.807677+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	d4af5efc60de02a777ba15e434490e826229f8f89a7ea7e9cbb24b6b701c5b9f
rollouts/028ac44d-71c0-4be5-a235-ad2e1697353d.json	59772625271f0d77ff4388147ef36b4d86114ffee3ce00e23c354c3c5c8635a7
rollouts/18e71903-405f-4fdd-9e86-d26c7999d977.json	04a36c4325b05a0fe29b755ab1ff96fc55ede7886f21e62782f0e1b73b70b523
rollouts/2cc0a656-e29c-4d64-8525-f42386d9b497.json	cae7e2e0ed54eb27a4c168e2ffbdb5d6e72d80eb394dd57ab0c911b618afdb80
rollouts/2d1074f7-1e2b-4b54-a7f5-75932f3edd34.json	736bd4b86617bd2216fd1c51920ab3bae729d10c855f3eb39d5bb3be42f8bd64
rollouts/31f2d57e-85b9-47ed-8440-29acc6ea489b.json	20de4851181645f0b37f37a2b315f8bf0c18436363b9d36f65eb38a2828ac314
rollouts/358e1aa0-da0a-4421-a6f5-67f4ecc59dd6.json	9f27c1560765cdaab2bac95af5694befb78fcee9b49cdc093f400c64a9475ae9
rollouts/7c8b66f3-3ce4-410d-b1a8-6387f5c07369.json	6dcbb97384142c168356c83ecabdcfdab69689cd0c14c8b5ab9c7e6c50ee62f6
rollouts/a3129789-0993-492e-8df4-88f31b25cf5d.json	c8fe326a667e06893fbb2f64f7314aaf98fbdcd49fa34f51a6da2498192d2cec
rollouts/c2ddef0b-20d0-4603-92f4-38e6af50c4ef.json	06c29f5f5da6bd9c3befd6462e527da11e85b98b9a4216aa0023639fbf59abeb
rollouts/c3604498-c12c-42a1-bdf2-8ba54361ef3a.json	ec554fb89c6e54a3a714497cba6009dc4411cdd257c3bc4a8bbbed6a11c39612c
spec	a746efd895becd9899c0efa181591730ab1a0e05d4cffa05e13559111138cc02
summary	ee6cff25b8f29ce86459ed40a4b98416b789603865edeadf541bff942b892158

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.

7c8b66f3 · 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.

c3604498 · cli:cli/opencode-muse-spark · seed 1 · 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.

2d1074f7 · cli:cli/opencode-muse-spark · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

18e71903 · cli:cli/opencode-muse-spark · seed 3 · 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.

a3129789 · cli:cli/opencode-muse-spark · seed 4 · 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.

028ac44d · cli:cli/opencode-muse-spark · seed 5 · 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.

c2ddef0b · cli:cli/opencode-muse-spark · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (30 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

2cc0a656 · cli:cli/opencode-muse-spark · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74

STEP	TOOL	PATH	RESULT	CHARS
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

358e1aa0 · cli:cli/opencode-muse-spark · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

31f2d57e · cli:cli/opencode-muse-spark · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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 Evidence

Every number above is recomputable from the sealed records below; each record verifies itself with `vvdex-env records verify`, and this document was regenerated from them with `vvdex-env records campaign`.

- `vvdex.annotation.image-object-1` → `fr-20260907-d8051665` · `recordDigest` `533b6d1f4d350c544088d42f0a3f3f3943d6e17c1095ac620bfe52406a8aeb5b`
- `vvdex.annotation.structured-data-1` → `fr-20260907-4b1e97bc` · `recordDigest` `72451cce4189c11abdcbbbe77e3dfc4c37f13375d9be7ea006984415d15d5bbc`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-c308bc27` · `recordDigest` `89f0a9f902db85a59371bfb233988ed854547dc96e9c1a2a8d6734bbf2943960`

13 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.