VVDEX FORGE

CERTIFIED EVALUATION CAMPAIGN

Muse Spark 1.3 native video via OpenRouter — twelve attempts, MP4 re-encode with exact frames

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

CURRENT

CAMPAIGN REPORT

2026-09-07

Evaluation scope:

meta/muse-spark-1.3

Prepared by:

VVDex Forge — regenerated from verified sealed records only

`fc-89575ed5687d` · 2 record(s) · every value recomputable

Contents

01 Executive summary

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

FIXED TASKS	SEALED RECORDS	MODEL LANES	ATTEMPTS
1	2	1	12

GRADED ATTEMPTS	CERTIFIED PASSES	LANE ERRORS (NOT GRADED)	WITHHELD / INVALID
10	4	2	0

Purpose

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

Annotation campaign scope. This campaign evaluates one fixed annotation tasks with two sealed attempt records on each task (2 sealed attempts in total). The sealed tasks cover video inputs. Video was evaluated through the sampled still frames recorded in every video attempt; these results do not establish native-video capability. Audio is N/A because no audio task appears in this campaign's sealed records. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

4

certified pass(es) across 10 graded attempts (12 attempts, 2 lane error(s) not graded) in 1 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-89575ed5687d · generated from 2 verified sealed record(s)

Key findings across the campaign

- `robot-video-1 · meta/muse-spark-1.3` · **ATTENTION** — 4 certified passes / 10 graded attempts; 6 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION VVDEX.ANNOTATION.ROBOT-VIDEO-1
meta/muse-spark-1.3	4/10 + 2 lane errors

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 one 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.robot-video-1`	A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

04 Campaign aggregate across this one exam

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
meta/muse-spark-1.3	**4**	10	2	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 one 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.

EXAM	LANE	SEED	EXCLUSION	RECORDED REASON
robot-video-1	meta/muse-spark-1.3	4	lane error	OpenRouter request cap reached (13 per rollout)
robot-video-1	meta/muse-spark-1.3	8	lane error	OpenRouter request cap reached (13 per rollout)

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
robot-video-1	`answer_key`	**`Applied consistently to 2 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 2 · robot-video-1

`vvdex.annotation.robot-video-1 · record fr-20260907-01cca689 · record digest b176309e60214fcf… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	1 of 2 graded rollouts	2 of 2 graded rollouts that called submit	0 of 2 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam `vvdex.annotation.robot-video-1` (v0.2.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 2 graded rollouts passed the independent hidden tests and were certified

1 of 2 graded rollouts passed. meta/muse-spark-1.3 completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 2 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was `submitted_failed` — submitted annotations rejected by the independent grader.

Key findings

F-01	meta/muse-spark-1.3 passed the independent hidden tests
F-02	submitted_failed × 1 (meta/muse-spark-1.3)
F-03	Statistical floor — 2 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 510.7s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 2 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.robot-video-1	0.2.0	annotation	2 (2 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

`0a0a51360cde69737c9b5caef895b8b7c2a5e2d66da89617d8776ac724b019b9`

EXAM FINGERPRINT

`85e026007af17db692614084…`

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

Evaluation configuration

REPORT	fr-20260907-01cca689
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.2.0
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 21:19:43	Exam evidence sealed (before any model saw the exam)
2026-09-07 09:09:41	Evaluation run started
2026-09-07 09:18:12	Evaluation run finished
2026-09-07 09:18:12	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
meta/muse-spark-1.3	openrouter	VVDex-configured	VVDex free-quota

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

Limits

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

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	http-429	0	list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 2. Graded (evaluable) rollouts: 2. 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
meta/muse-spark-1.3 `meta/muse-spark-1.3`	2	2	1 / 2	1	0	2 / 2	255.1s	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@2
meta/muse-spark-1.3	2	2	0	1	50.0%	[9%, 91%]	0, 1	50.0%	100.0%

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1, 2; 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	OPENROUTER:META/MUSE-SPARK-1.3
Inspect	2 / 2
Annotate	2 / 2
Validate	2 / 2
Submit	2 / 2
Graded	2 / 2
Passed	1 / 2

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
openrouter:meta/muse-spark-1.3	Passed	Passed	1 / 2	0

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

Annotation evidence

Modality: video. **Rubric:** 1.1.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission ae21e559-693d-4e31-a388-65074b74ce97

Score	Items	Valid / invalid	Types
0.6632	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.5729
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	0.5	eventPrecision	0.5	eventRecall	0.5
eventTemporalIoU	0.4667	matchedCount	6	mismatchCount	5
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	0.9	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 7a9d5c84-a8e4-4e4e-8284-3632138829dd

Score	Items	Valid / invalid	Types
0.8703	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.7359
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.875	matchedCount	6	mismatchCount	3
segmentF1	1	segmentPrecision	1	segmentRecall	1

segmentTemporalIoU	1	Track consistency on matched boxes	1

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

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

Quality axes

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

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
boxPrecision	recorded	N/A	1	–	–	
boxRecall	recorded	N/A	1	–	–	
boxF1	recorded	N/A	1	–	–	
boxIoU	recorded	N/A	0.6544	–	–	
classificationPrecision	recorded	N/A	1	–	–	
classificationRecall	recorded	N/A	1	–	–	
classificationF1	recorded	N/A	1	–	–	
classificationAccuracy	recorded	N/A	1	–	–	
eventPrecision	recorded	N/A	0.75	–	–	
eventRecall	recorded	N/A	0.75	–	–	
eventF1	recorded	N/A	0.75	–	–	
eventTemporalIoU	recorded	N/A	0.6708	–	–	
segmentPrecision	recorded	N/A	1	–	–	
segmentRecall	recorded	N/A	1	–	–	
segmentF1	recorded	N/A	1	–	–	
segmentTemporalIoU	recorded	N/A	0.95	–	–	
trackingConsistency	recorded	N/A	1	–	–	

QUALITY SCORE (MEAN)

0.7667

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
openrouter:meta/muse-spark-1.3	2	9 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

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

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
openrouter:meta/muse-spark-1.3	2	0	none	0 / 2	0 / 2	0.0%	0

No grounding failures and no false pass claims were detected across these 2 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

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

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
openrouter:meta/muse-spark-1.3	2	231 126	115 563	not billed	1 / 2	50.0%

Findings

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

F-01 PASS

meta/muse-spark-1.3 passed the independent hidden tests

Severity: PASS | Evidence: Model outcomes · Stage funnel · Appendix A

1 of 2 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 1 (meta/muse-spark-1.3)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted annotations rejected by the independent grader.

F-02 · derived from the recorded run data

F-03 NOTE

Statistical floor — 2 rollout(s) is below the 10-rollout rate floor

Severity: NOTE | Evidence: Appendix D

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

F-03 · derived from the recorded run data

Recommendations

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

R–01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

meta/muse-spark-1.3

WHAT WE WOULD LOOK AT FIRST

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

R–02

For anyone reading the percentages

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

Verification

RUN ID

live-20260907T090941Z

EXAM FINGERPRINT

85e026007af17db692614084f12ff40c5bbf2348bcfda0a01deb95aad8928f4c

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

b176309e60214fcf8b63317ca97122f228fc235e4bd09ad5dd9083b2795b6906

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

SOURCE RECORD FILE

4683299eba4f1559ac3de97ca68d2537f8c27b61a1a313339c70a6f31980fead

sha256 of fr-20260907-01cca689.json as written beside the standalone report.

SOURCE STANDALONE REPORT

67e6be92b5db13c99f067bf99372c05b996f0fd7a7f62d668a5263f0ac4423d6

sha256 of fr-20260907-01cca689.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

dd2cccdb36024b961dfabaabbfd73350749d7326b63e3d80b28936ec2c8b65c6

sha256 of fr-20260907-01cca689.pdf.

RUN EVIDENCE BUNDLE

01cca6890a7e0b3a93c94256ccd921c27f4258ef23f0b293747b3c6cbcc1f605

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

CERTIFIED EVIDENCE SEAL

0a0a51360cde69737c9b5caef895b8b7c2a5e2d66da89617d8776ac724b019b9

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

Measurement history

Previous run on this exam: `fr-20260907-023b5cdb`. Model progress over time on this exam is the difference between that record and this one — same frozen tree, same hidden grader, same limits. Anything else that moved between them moved outside this exam.

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

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
ae21e559	meta/muse-spark-1.3	0	submitted_failed	yes	Graded	fail	3	132 488	297.8s
7a9d5c84	meta/muse-spark-1.3	1	submitted_passed	yes	Passed	pass	2	98 638	212.4s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
01cca6890a7e0b3a93c94256ccd921c27f4258ef23f0b293747b3c6cbcc1f605	live-20260907T090941Z	2026-09-07T09:09:41.755354+00:00	2026-09-07T09:18:12.493652+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	0819a4d090cc57d82b845d97b5938a46ee8c066a4e56645b2673abc8639f24df
rollouts/7a9d5c84-a8e4-4e4e-8284-3632138829dd.json	3d26a618937673d6ef0cab7ab213616fcb990040466e7ce21d07e0d374a9406f
rollouts/ae21e559-693d-4e31-a388-65074b74ce97.json	772d29eb46dcf5f0ad22d150a9652019b26e765533f49c004bd3520719947ce7
spec	4339f4f6d326675825048a4d2aa9272ec26a72dde02a40652bddfcbc8ec834b4
summary	0be2db2ed0bdc91e0698318f2a26b3f4befdf62d73529f76e16f70e246a93bbe

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.

ae21e559 · openrouter:meta/muse-spark-1.3 · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	write_file	submission.json	ok	64
8	run_tests	—	ok	160
9	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.

7a9d5c84 · openrouter:meta/muse-spark-1.3 · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
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 2 of 2 · robot-video-1

`vvdex.annotation.robot-video-1 · record fr-20260907-023b5cdb · record digest 40a633cc2fd5e56e… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
1	3 of 8 graded rollouts	8 of 8 graded rollouts that called submit	2 of 10 attempts

Purpose

This report presents a controlled evaluation of 1 model lane(s) against the certified exam vvdex.annotation.robot-video-1 (v0.2.0, 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

3 of 8

graded rollouts passed the independent hidden tests and were certified — 2 of 10 attempts not graded (lane errors)

3 of 8 graded rollouts passed. meta/muse-spark-1.3 completed the applicable annotation workflow and were accepted by the independent annotation grader. The most common way to fail was `submitted_failed` — submitted annotations rejected by the independent grader.

Key findings

F-01	meta/muse-spark-1.3 passed the independent hidden tests
F-02	submitted_failed × 5 (meta/muse-spark-1.3)
F-03	model_api_failure × 2 (meta/muse-spark-1.3)

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: 2824.7s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s 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.robot-video-1` was certified before this run: its reference solution passes, an empty submission fails, its grader distinguishes near-misses, and its sealed evidence is unchanged by this evaluation (see Certification evidence).

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.robot-video-1	0.2.0	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

`0a0a51360cde69737c9b5caef895b8b7c2a5e2d66da89617d8776ac724b019b9`

EXAM FINGERPRINT

`85e026007af17db692614084…`

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

Evaluation configuration

REPORT	fr-20260907-023b5cdb
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.robot-video-1 · v0.2.0
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 21:19:43	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:22:33	Evaluation run started
2026-09-07 09:09:38	Evaluation run finished
2026-09-07 09:09:38	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
meta/muse-spark-1.3	openrouter	VVDex-configured	VVDex free-quota

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

Limits

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

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	http-429	0	list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 10. Graded (evaluable) rollouts: 8. Not graded — lane errors: 2. 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
meta/muse-spark-1.3 `meta/muse-spark-1.3`	10	8	3 / 8	5	2	8 / 8	267.1s	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@8
meta/muse-spark-1.3	10	8	2	3	37.5%	[14%, 69%]	0–9	37.5%	82.1%	98.2%	100.0%

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1, 3, 5, 8; 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	OPENROUTER:META/MUSE-SPARK-1.3
Inspect	8 / 8
Annotate	8 / 8
Validate	8 / 8
Submit	8 / 8
Graded	8 / 8
Passed	3 / 8

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
openrouter:meta/muse-spark-1.3	Passed	Graded	3 / 8	2

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

Annotation evidence

Modality: video. **Rubric:** 1.1.

A robot-catching video is reviewed for temporal events, frame-level object boxes, track identity and occlusion.

Submission dd120ba7-a5d4-42fd-8c4a-7c312d18a602

Score	Items	Valid / invalid	Types
0.7638	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.579
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.7125	matchedCount	6	mismatchCount	4
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	1	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission cc1b55f2-6d54-436d-8737-aa81aae3a220

Score	Items	Valid / invalid	Types
0.7003	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	0.5	boxIoU	0.5528
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.5982	matchedCount	6	mismatchCount	5
segmentF1	1	segmentPrecision	1	segmentRecall	1

segmentTemporalIoU	0.9	Track consistency on matched boxes	1

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 415869af-1240-4a10-bf95-d12fccfcb96c

Score	Items	Valid / invalid	Types
0.7802	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	0	boxIoU	0.4656
boxPrecision	0	boxRecall	0	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.875	matchedCount	6	mismatchCount	3
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	1	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 8fe2d8d3-b887-4b2f-ae75-a28eda0bb8be

Score	Items	Valid / invalid	Types
0.6654	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.5797
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1

eventF1	0.5	eventPrecision	0.5	eventRecall	0.5
eventTemporalIoU	0.4667	matchedCount	6	mismatchCount	5
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	0.9	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 0dbb95df-4a7f-438d-b3c3-74a1e68017eb

Score	Items	Valid / invalid	Types
0.7908	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	0.5	boxIoU	0.4975
boxPrecision	0.5	boxRecall	0.5	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.875	matchedCount	6	mismatchCount	3
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	1	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 078d6f61-7dd8-4c2a-8f74-eb4049a5b391

Score	Items	Valid / invalid	Types
0.8217	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.5902
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.875	matchedCount	6	mismatchCount	3
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	1	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission 2efda1c5-b89e-4014-a4cf-161e5c35599f

Score	Items	Valid / invalid	Types
0.6826	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	0	boxIoU	0.459
boxPrecision	0	boxRecall	0	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.6125	matchedCount	6	mismatchCount	5
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	0.9524	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

Submission be8a7b18-8bd0-4ebd-9588-0211ca846bc1

Score	Items	Valid / invalid	Types
0.7473	6	6 / 0	box, classification, event, segment

Measured metrics

annotationCount	6	boxF1	1	boxIoU	0.5295
boxPrecision	1	boxRecall	1	classificationAccuracy	1
classificationF1	1	classificationPrecision	1	classificationRecall	1
eventF1	1	eventPrecision	1	eventRecall	1
eventTemporalIoU	0.7125	matchedCount	6	mismatchCount	4
segmentF1	1	segmentPrecision	1	segmentRecall	1
segmentTemporalIoU	1	Track consistency on matched boxes	1		

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

Validation error codes

None recorded

Disagreement

Agreement review	Not measured

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

Quality axes

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

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
boxPrecision	recorded	N/A	0.625	–	–	
boxRecall	recorded	N/A	0.625	–	–	
boxF1	recorded	N/A	0.625	–	–	
boxIoU	recorded	N/A	0.5317	–	–	
classificationPrecision	recorded	N/A	1	–	–	
classificationRecall	recorded	N/A	1	–	–	
classificationF1	recorded	N/A	1	–	–	
classificationAccuracy	recorded	N/A	1	–	–	
eventPrecision	recorded	N/A	0.9375	–	–	
eventRecall	recorded	N/A	0.9375	–	–	
eventF1	recorded	N/A	0.9375	–	–	
eventTemporalIoU	recorded	N/A	0.7159	–	–	
segmentPrecision	recorded	N/A	1	–	–	
segmentRecall	recorded	N/A	1	–	–	
segmentF1	recorded	N/A	1	–	–	
segmentTemporalIoU	recorded	N/A	0.969	–	–	
trackingConsistency	recorded	N/A	1	–	–	

QUALITY SCORE (MEAN)

0.744

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
openrouter:meta/muse-spark-1.3	10	10 / 24	0	0 / 0	0	0	4	1

Each column is defined in Appendix D.

Hallucination & overclaim

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

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
openrouter:meta/muse-spark-1.3	10	0	none	0 / 10	0 / 10	0.0%	4

No grounding failures and no false pass claims were detected across these 10 rollout(s). openrouter:meta/muse-spark-1.3 re-read previously returned material 4 time(s); other lanes did not.

Every term above is defined in Appendix D.

Efficiency & telemetry

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

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
openrouter:meta/muse-spark-1.3	8	1 073 893	134 236	not billed	3 / 8	37.5%

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

meta/muse-spark-1.3 passed the independent hidden tests

Severity: PASS | Evidence: Model outcomes · Stage funnel · Appendix A

3 of 8 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 5 (meta/muse-spark-1.3)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted annotations rejected by the independent grader.

F-02 · derived from the recorded run data

F-03 LANE / INFRA

model_api_failure × 2 (meta/muse-spark-1.3)

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

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

F-03 · derived from the recorded run data

Recommendations

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

R–01

PROBLEM

Submitted annotations rejected by the independent grader.

AFFECTED LANES

meta/muse-spark-1.3

WHAT WE WOULD LOOK AT FIRST

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

R–02

PROBLEM

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

AFFECTED LANES

meta/muse-spark-1.3

WHAT WE WOULD LOOK AT FIRST

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

Verification

RUN ID

live-20260907T082233Z

EXAM FINGERPRINT

85e026007af17db692614084f12ff40c5bbf2348bcfda0a01deb95aad8928f4c

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

40a633cc2fd5e56e5a165ae5e4f8d8b27ceb1278fdf4ee38e46267b106cc8655

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

SOURCE RECORD FILE

079d54cc556ec585dc806c66514f025eca04c9325bf64854e9421669b3f09056

sha256 of fr-20260907-023b5cdb.json as written beside the standalone report.

SOURCE STANDALONE REPORT

859d36b8815533e2c372f26d1e9a5c5ab9469ee9223d91aff31bf11e0981f5cc

sha256 of fr-20260907-023b5cdb.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

1ec463793e9db2b58d767b052d076e47478edf42892d7f27adb687270f7a44a9

sha256 of fr-20260907-023b5cdb.pdf.

RUN EVIDENCE BUNDLE

023b5cdb0bec7cdb5402199e372974bbbaacea8cfea639dc9fbf118bbcef9d1b

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

CERTIFIED EVIDENCE SEAL

0a0a51360cde69737c9b5caef895b8b7c2a5e2d66da89617d8776ac724b019b9

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
dd120ba7	meta/muse-spark-1.3	0	submitted_passed	yes	Passed	pass	3	129 862	279.4s
cc1b55f2	meta/muse-spark-1.3	1	submitted_failed	yes	Graded	fail	4	178 087	333.7s
415869af	meta/muse-spark-1.3	2	submitted_failed	yes	Graded	fail	3	133 056	244.8s
8fe2d8d3	meta/muse-spark-1.3	3	submitted_failed	yes	Graded	fail	3	172 970	290.5s
da93e045	meta/muse-spark-1.3	4	model_api_failure	no	Not graded — lane error	not graded	6	n/a	373.3s
0dbb95df	meta/muse-spark-1.3	5	submitted_failed	yes	Graded	fail	2	102 569	250.4s
078d6f61	meta/muse-spark-1.3	6	submitted_passed	yes	Passed	pass	2	112 954	213.4s
2efda1c5	meta/muse-spark-1.3	7	submitted_failed	yes	Graded	fail	3	129 912	254.6s
1eb52e4b	meta/muse-spark-1.3	8	model_api_failure	no	Not graded — lane error	not graded	5	n/a	311.8s
be8a7b18	meta/muse-spark-1.3	9	submitted_passed	yes	Passed	pass	3	114 483	270.3s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	5	Submitted annotations rejected by the independent grader.
model_api_failure	2	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
023b5cdb0bec7cdb5402199e372974bbbaacea8cfea639dc9fbf118bbcef9d1b	live-20260907T082233Z	2026-09-07T08:22:33.271069+00:00	2026-09-07T09:09:38.009136+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	c50ed23ef6b4c129b4232edc2fd8661c684e14943f9775b07eb60fdf1235beda
rollouts/078d6f61-7dd8-4c2a-8f74-eb4049a5b391.json	21d1d4ac0afff1bdcd7f1ece175225fc0b05e3f29a9263ab785ac8d7c57f6e45
rollouts/0dbb95df-4a7f-438d-b3c3-74a1e68017eb.json	2d255e71772a144234665a7fb3b8d7c1dce99fdc37f7827b56493b04d13f1159
rollouts/1eb52e4b-3662-4fe9-80e5-445d7222a8cf.json	efa0138630ce73d80c14d0913a32003d43cb3d5bfce89d90f0d66b4f0b32ae96
rollouts/2efda1c5-b89e-4014-a4cf-161e5c35599f.json	92a6ccd12bff499e88854d3cb385d6f23cd9d51cc5616759c239389cb7221112
rollouts/415869af-1240-4a10-bf95-d12fccfcb96c.json	a9ab8a3ebfaf30005278931ed8b3750b765989fa5fc38ac9a1f4e57a36bbacd0
rollouts/8fe2d8d3-b887-4b2f-ae75-a28eda0bb8be.json	5e7ff9c2e01d5e1410ad76fbdcce0a9ad12f3d78db7414e8c79374850f05c528
rollouts/be8a7b18-8bd0-4ebd-9588-0211ca846bc1.json	4f4731b37f62645d23028f5e25536acbbd43fd281337aa2eb0d9ba7641090765
rollouts/cc1b55f2-6d54-436d-8737-aa81aae3a220.json	ee29ada9d5ba47c6b5f7c0424aab712f492b09a545a09bfe8b3de65f9bfe1b74
rollouts/da93e045-0855-4ec9-a6fc-5ea7a6963d22.json	6a0deba95998a940dff94c35ae0838f458b4f707c149033078d8e2cc12ba5c99
rollouts/dd120ba7-a5d4-42fd-8c4a-7c312d18a602.json	79727bf726edd7458694df404ffac925fba97bf93ea83134ae9c24e68390ccfb
spec	13a78198f4184d167af717ac09b899f9c9b07899b2a73f4f0c68fb4f4005ed6e
summary	5db612d49a6f66d4f0ee16b77f60e86f6c1f4df368ddd0dc1fe7f05ff01a7c42

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.

dd120ba7 · openrouter:meta/muse-spark-1.3 · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	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.

cc1b55f2 · openrouter:meta/muse-spark-1.3 · seed 1 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	list_files	assets	ok	17
9	edit_file	submission.json	ok	53
10	run_tests	—	ok	160

STEP	TOOL	PATH	RESULT	CHARS
11	run_tests	—	ok	160
12	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.

415869af · openrouter:meta/muse-spark-1.3 · seed 2 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	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.

8fe2d8d3 · openrouter:meta/muse-spark-1.3 · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	1 119
1	list_files	.	ok	79
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	failed	29
6	read_file	submission.json	ok	711
7	edit_file	submission.json	failed	29
8	write_file	submission.json	ok	64
9	run_tests	—	ok	160
10	write_file	submission.json	ok	64
11	run_tests	—	ok	160
12	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.

da93e045 · openrouter:meta/muse-spark-1.3 · seed 4 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	160
10	edit_file	submission.json	ok	53
11	read_file	submission.json	ok	709
12	edit_file	submission.json	ok	53

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.

0dbb95df · openrouter:meta/muse-spark-1.3 · seed 5 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
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.

078d6f61 · openrouter:meta/muse-spark-1.3 · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	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.

2efda1c5 · openrouter:meta/muse-spark-1.3 · seed 7 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	edit_file	submission.json	ok	53
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	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.

1eb52e4b · openrouter:meta/muse-spark-1.3 · seed 8 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64

STEP	TOOL	PATH	RESULT	CHARS
5	edit_file	submission.json	failed	29
6	read_file	submission.json	ok	629
7	edit_file	submission.json	ok	53
8	edit_file	submission.json	ok	53
9	edit_file	submission.json	ok	53
10	read_file	submission.json	ok	629
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	160

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.

be8a7b18 · openrouter:meta/muse-spark-1.3 · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	79
1	read_file	TASK.md	ok	1 119
2	read_file	annotation-task.json	ok	4 867
3	read_file	submission.json	ok	74
4	write_file	submission.json	ok	64
5	edit_file	submission.json	ok	53
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	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 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.robot-video-1 → fr-20260907-01cca689 · recordDigest b176309e60214fcf8b63317ca97122f228fc235e4bd09ad5dd9083b2795b6906`
- `vvdex.annotation.robot-video-1 → fr-20260907-023b5cdb · recordDigest 40a633cc2fd5e56e5a165ae5e4f8d8b27ceb1278fdf4ee38e46267b106cc8655`

12 Attestation

Certified results appear only where evidence proves them. Withheld, invalid and lane-error cells are never presented as model results, and no historical evidence was modified in producing this document. Counts are exact; below ten rollouts per cell no rate or ranking beyond the counts is asserted.