

Free API lanes — text and structured, ten attempts each

2 fixed tasks · 4 sealed attempt records · 124
graded attempts · 67 certified passes

CURRENT
CAMPAIGN REPORT
2026-09-07

Evaluation scope:
openai/gpt-oss-20b
openai/gpt-oss-120b
@cf/openai/gpt-oss-20b
codestral-latest
mistral-code-latest
ministral-14b-latest
ministral-8b-latest

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

fc-3377c49ce724 · 4 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	5
03.	Environment families	6
04.	Campaign aggregate across these two exams	7
05.	Harness status	8
06.	Cells excluded from capability	9
07.	What the outcomes mean	10
08.	What this report withholds, and why	11
09.	Attempt record 1 of 2 · structured-data-1	12
10.	Attempt record 2 of 2 · structured-data-1	52
11.	Attempt record 1 of 2 · text-labeling-1	103
12.	Attempt record 2 of 2 · text-labeling-1	154
13.	Evidence	186
14.	Attestation	187

01 Executive summary

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

FIXED TASKS 2	SEALED RECORDS 4	MODEL LANES 7	ATTEMPTS 136
GRADED ATTEMPTS 124	CERTIFIED PASSES 67	LANE ERRORS (NOT GRADED) 12	WITHHELD / INVALID 0

Purpose

This report presents a controlled evaluation of 7 model lane(s) against 2 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 two fixed annotation tasks with two sealed attempt records on each task (4 sealed attempts in total). The sealed tasks cover structured, text inputs. Audio is N/A because no audio task appears in this campaign's sealed records. Quality measurements are reported per task and remain separate from the strict certified-pass gate.

Outcome

67

certified pass(es) across 124 graded attempts (136 attempts, 12 lane error(s) not graded) in 14 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-3377c49ce724 · generated from 4 verified sealed record(s)

Key findings across the campaign

- structured-data-1 · openai/gpt-oss-20b · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · openai/gpt-oss-120b · **ATTENTION** — 7 certified passes / 10 graded attempts; 3 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · @cf/openai/gpt-oss-20b · **ATTENTION** — 9 certified passes / 10 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · codestral-latest · **ATTENTION** — 9 certified passes / 10 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · mistral-code-latest · **ATTENTION** — 9 certified passes / 10 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · ministral-14b-latest · **ATTENTION** — 8 certified passes / 10 graded attempts; 2 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

- structured-data-1 · ministral-8b-latest · **ATTENTION** — 1 certified passes / 10 graded attempts; 9 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · openai/gpt-oss-20b · **ATTENTION** — 1 certified passes / 1 graded attempts; 0 model failures; 9 excluded (9 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · openai/gpt-oss-120b · **ATTENTION** — 3 certified passes / 3 graded attempts; 0 model failures; 3 excluded (3 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · @cf/openai/gpt-oss-20b · **PASS** — 10 certified passes / 10 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · codestral-latest · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · mistral-code-latest · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · ministral-14b-latest · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · ministral-8b-latest · **ATTENTION** — 0 certified passes / 10 graded attempts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION · STRUCTURED-DATA-1 VVDEX.ANNOTATION.STRUCTURED-DATA-1	ANNOTATION · TEXT-LABELING-1 VVDEX.ANNOTATION.TEXT-LABELING-1
@cf/openai/gpt-oss-20b	9/10	10/10
openai/gpt-oss-20b	10/10	1/1 + 9 lane errors
openai/gpt-oss-120b	7/10	3/3 + 3 lane errors
codestral-latest	9/10	0/10
mistral-code-latest	9/10	0/10
ministral-14b-latest	8/10	0/10
ministral-8b-latest	1/10	0/10

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

03 Environment families

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

FAMILY	EXAM	WHAT IT TESTS
Multimodal annotation	vvdex.annotation.structured-data-1	Invented tabular records are reviewed for normalization, validity, duplicates, anomalies and relationships.
Multimodal annotation	vvdex.annotation.text-labeling-1	An invented message is reviewed for classification and entity spans with consistent offsets.

04 Campaign aggregate across these two exams

MODEL LANE	CERTIFIED PASSES	GRADED ATTEMPTS	LANE ERRORS	WITHHELD	INVALID
@cf/openai/gpt-oss-20b	19	20	0	0	0
openai/gpt-oss-20b	11	11	9	0	0
openai/gpt-oss-120b	10	13	3	0	0
codestral-latest	9	20	0	0	0
mistral-code-latest	9	20	0	0	0
ministral-14b-latest	8	20	0	0	0
ministral-8b-latest	1	20	0	0	0

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

05 Harness status

1 of 2 exams include harness-suspect records (text-labeling-1). Each fired rule is classified below; a verdict-bearing rule the record cannot clear from its own evidence withholds the affected cells. **0 cell(s) withheld** on that basis.

Here R1 records provider/API interruption as a diagnostic; it does not assert a grader defect or turn an ungraded lane error into a model result.

text-labeling-1 · R1 · diagnostic

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

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
text-labeling-1	openai/gpt-oss-120b	3	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): ...
text-labeling-1	openai/gpt-oss-120b	4	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): ...
text-labeling-1	openai/gpt-oss-120b	5	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): ...
text-labeling-1	openai/gpt-oss-20b	0	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	1	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	3	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	4	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	5	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	6	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	7	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	8	lane error	HTTP 400: Failed to parse tool call arguments as JSON
text-labeling-1	openai/gpt-oss-20b	9	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-20b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): ...

07 What the outcomes mean

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

08 What this report withholds, and why

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

EXAM	DISCLOSURE CLASS	RULE APPLIED
structured-data-1	answer_key	Applied consistently to 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.
text-labeling-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 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260907-811ca1e7 · record digest 50cbe9a820d64927... · harness clean · containment clean

Executive summary

MODELS TESTED 3	CERTIFIED PASSES 26 of 30 graded rollouts	SUBMITTED 30 of 30 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 30 attempts
---------------------------------------	--	--	--

Purpose

This report presents a controlled evaluation of 3 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

26 of 30

graded rollouts passed the independent hidden tests and were certified

26 of 30 graded rollouts passed. @cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b 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 @cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests

F-02 submitted_failed × 4 (@cf/openai/gpt-oss-20b, openai/gpt-oss-120b)

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: 2173.0s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 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.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	30 (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-20260907-811ca1e7
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.structured-data-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:39:35	Evaluation run started
2026-09-07 09:15:48	Evaluation run finished
2026-09-07 09:15:48	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
openai/gpt-oss-20b	groq	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota
@cf/openai/gpt-oss-20b	cloudflare	VVDex-configured	VVDex free-quota

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 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY http-429	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: 30. Graded (evaluable) rollouts: 30. 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
openai/gpt-oss-20b openai/gpt-oss-20b	10	10	10 / 10	0	0	10 / 10	62.5s	Passed
openai/gpt-oss-120b openai/gpt-oss-120b	10	10	7 / 10	3	0	10 / 10	77.0s	Passed
@cf/openai/gpt-oss-20b @cf/openai/gpt-oss-20b	10	10	9 / 10	1	0	10 / 10	19.0s	Passed

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
openai/gpt-oss-20b	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
openai/gpt-oss-120b	10	10	0	7	70.0%	[40%, 89%]	0–9	70.0%	99.2%	100.0%	100.0%
@cf/openai/gpt-oss-20b	10	10	0	9	90.0%	[60%, 98%]	0–9	90.0%	100.0%	100.0%	100.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
openai/gpt-oss-20b	100.0%	[72%, 100%]	124 318	not billed	62.5s	0.0%	none
openai/gpt-oss-120b	70.0%	[40%, 89%]	165 189	not billed	77.0s	0.0%	submitted_failed
@cf/openai/gpt-oss-20b	90.0%	[60%, 98%]	134 012	not billed	19.0s	0.0%	submitted_failed

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

This is not a leaderboard. Not separated (intervals overlap): openai/gpt-oss-20b vs openai/gpt-oss-120b; openai/gpt-oss-20b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-120b vs @cf/openai/gpt-oss-20b. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	GROQ:OPENAI/GPT-OSS-20B	GROQ:OPENAI/GPT-OSS-120B	CLOUDFLARE:@CF/OPENAI/GPT-OSS-20B
Inspect	10 / 10	10 / 10	10 / 10
Annotate	10 / 10	10 / 10	10 / 10
Validate	10 / 10	10 / 10	10 / 10
Submit	10 / 10	10 / 10	10 / 10
Graded	10 / 10	10 / 10	10 / 10
Passed	10 / 10	7 / 10	9 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
groq:openai/gpt-oss-20b	Passed	Passed	10 / 10	0
groq:openai/gpt-oss-120b	Passed	Passed	7 / 10	0
cloudflare:@cf/openai/gpt-oss-20b	Passed	Passed	9 / 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 b99939d9-d884-423e-8d6d-e7dce1134325

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 79a654b1-bdb8-42f9-9c4b-20dd953c5627

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 769d77be-3d37-454d-aa1c-1b65c80e77ff

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 e650c011-9065-4a93-961b-1e4dab65ba55

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 ae7a3122-fd75-4b14-b8f5-fd6069493b36

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 d8479248-5aa4-41f2-8773-a16295e57890

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 6a3337e7-1d52-493e-b53a-8a6f30257d35

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 2da20181-72a7-4832-ac72-cb549fc99df9

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 97a444e5-8609-4a78-b30a-23a72f8fa5bb

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 8e22d41b-ccf9-4a19-a4b6-16caf07741d1

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 f655f3c1-963f-47d4-8835-605b93026762

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 8d664dab-d3e1-43af-b30a-765f9beb9135

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 6061e717-ae4b-466f-9b21-ded38e70c8c5

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 6ff0c5ed-b4c7-4e40-806d-bfb4a370c138

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 7b8cbba4-90c3-415e-b94d-6d780b42f8ba

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 21cc70ac-4c9e-4974-8dc7-55777ba722b3

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 5756f518-8d8e-497b-8bb3-9d98305faa66

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 a5066bc6-9497-4d5d-9f5f-732e79d9dcbf

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 b2db85b4-95ce-4917-9a9f-05ddf0e47ee4

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 0fb1ed25-19d4-434b-accf-9850cc322cd4

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 55c7a7d4-6651-4565-b585-75ee1e4d3a51

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 8f0b2bac-c328-4f5c-b28c-4283f200c96d

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 9f158050-2fba-4365-8f99-ea5b029a1338

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 77a79aab-2c86-4caa-91e7-975403752155

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 faa01cab-e8cf-46ae-900d-77237c7d9145

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 53f368f1-d8f0-4e27-91b8-f648abda7fc2

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 d2421f8c-0dfa-41d9-8f54-dbda8a1c5320

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 b30cd68a-718e-4b9e-b6f6-3763dfccb9d0

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 1f2c4f23-71e3-4f6d-a2c9-c83665971442

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 833bcaa3-8d47-48e5-a300-8aaedb7bccfd

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

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	0.9556	–	–	
rowRecall	recorded	N/A	0.9556	–	–	
rowF1	recorded	N/A	0.9556	–	–	
rowAccuracy	recorded	N/A	0.9556	–	–	

QUALITY SCORE (MEAN)

0.9667

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
groq:openai/gpt-oss-20b	10	7 / 24	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	10	8 / 24	0	0 / 0	0	0	1	0
cloudflare:@cf/openai/gpt-oss-20b	10	7 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

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

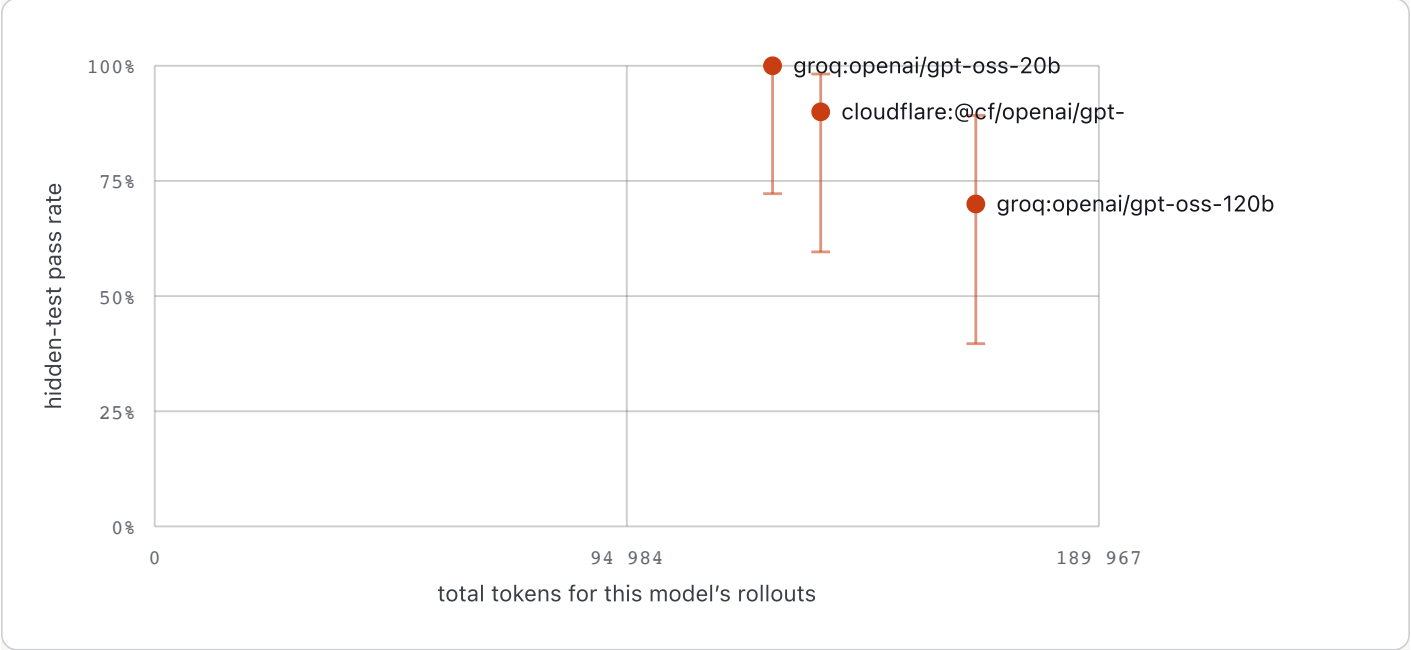
MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
groq:openai/gpt-oss-20b	10	0	none	0 / 10	0 / 10	0.0%	0
groq:openai/gpt-oss-120b	10	0	none	0 / 10	0 / 10	0.0%	1
cloudflare:@cf/openai/gpt-oss-20b	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 30 rollout(s). groq:openai/gpt-oss-120b re-read previously returned material 1 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.



Only lanes that report token telemetry are charted. The vertical bar through each point is that model's 95% interval, drawn to the same scale as the axis — where it spans most of the axis, the point is one measurement with a wide interval, not a position on a frontier.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
groq:openai/gpt-oss-20b	10	124 318	12 431	not billed	10 / 10	100.0%
groq:openai/gpt-oss-120b	10	165 189	16 518	not billed	7 / 10	70.0%
cloudflare:@cf/openai/gpt-oss-20b	10	134 012	13 401	not billed	9 / 10	90.0%

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

@cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

26 of 30 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 × 4 (@cf/openai/gpt-oss-20b, openai/gpt-oss-120b)

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

@cf/openai/gpt-oss-20b, openai/gpt-oss-120b

WHAT WE WOULD LOOK AT FIRST

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

Verification

RUN ID

live-20260907T083935Z

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

50cbe9a820d64927b58510039c580bbcd3bd161565d085363f841eb867181fab

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

SOURCE RECORD FILE

0b8fb9e6e3f0863c690ce46327edc5cfc2cbd12fc2be0bdc6484d3e38ec21d75

sha256 of fr-20260907-811ca1e7.json as written beside the standalone report.

SOURCE STANDALONE REPORT

5f22411a34bd090a856fc623775c664f7f1783eb54bb5fa31cc4fdfdcede6d74

sha256 of fr-20260907-811ca1e7.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

7758b66666b7082b98895b578e2c3283a64cbc3dfe1344477bd7d7b879c370da

sha256 of fr-20260907-811ca1e7.pdf.

RUN EVIDENCE BUNDLE

811ca1e736f76331d8125a2a1fd6972080e7f8d7dcdae610287199470d7896dd

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
b99939d9	openai/gpt-oss-20b	0	submitted_passed	yes	Passed	pass	1	10 152	12.1s
79a654b1	openai/gpt-oss-20b	1	submitted_passed	yes	Passed	pass	1	9 837	46.4s
769d77be	openai/gpt-oss-20b	2	submitted_passed	yes	Passed	pass	1	14 408	97.0s
e650c011	openai/gpt-oss-20b	3	submitted_passed	yes	Passed	pass	1	14 582	70.4s
ae7a3122	openai/gpt-oss-20b	4	submitted_passed	yes	Passed	pass	1	14 459	86.5s
d8479248	openai/gpt-oss-20b	5	submitted_passed	yes	Passed	pass	1	11 608	54.3s
6a3337e7	openai/gpt-oss-20b	6	submitted_passed	yes	Passed	pass	1	14 288	88.4s
2da20181	openai/gpt-oss-20b	7	submitted_passed	yes	Passed	pass	1	12 584	63.9s
97a444e5	openai/gpt-oss-20b	8	submitted_passed	yes	Passed	pass	1	12 412	64.2s
8e22d41b	openai/gpt-oss-20b	9	submitted_passed	yes	Passed	pass	1	9 988	42.1s
f655f3c1	openai/gpt-oss-120b	0	submitted_failed	yes	Graded	fail	1	13 975	50.0s
8d664dab	openai/gpt-oss-120b	1	submitted_passed	yes	Passed	pass	1	14 450	62.4s
6061e717	openai/gpt-oss-120b	2	submitted_passed	yes	Passed	pass	1	17 631	103.5s
6ff0c5ed	openai/gpt-oss-120b	3	submitted_failed	yes	Graded	fail	1	17 584	90.5s
7b8cbba4	openai/gpt-oss-120b	4	submitted_failed	yes	Graded	fail	1	17 680	87.5s
21cc70ac	openai/gpt-oss-120b	5	submitted_passed	yes	Passed	pass	1	18 060	80.3s
5756f518	openai/gpt-oss-120b	6	submitted_passed	yes	Passed	pass	1	14 997	68.8s
a5066bc6	openai/gpt-oss-120b	7	submitted_passed	yes	Passed	pass	1	14 253	66.9s
b2db85b4	openai/gpt-oss-120b	8	submitted_passed	yes	Passed	pass	1	14 032	48.4s
0fb1ed25	openai/gpt-oss-120b	9	submitted_passed	yes	Passed	pass	2	22 527	111.2s
55c7a7d4	@cf/openai/gpt-oss-20b	0	submitted_passed	yes	Passed	pass	1	16 045	22.0s
8f0b2bac	@cf/openai/gpt-oss-20b	1	submitted_passed	yes	Passed	pass	1	12 935	17.3s
9f158050	@cf/openai/gpt-oss-20b	2	submitted_passed	yes	Passed	pass	1	12 383	20.7s
77a79aab	@cf/openai/gpt-oss-20b	3	submitted_passed	yes	Passed	pass	1	12 611	19.0s
faa01cab	@cf/openai/gpt-oss-20b	4	submitted_passed	yes	Passed	pass	1	12 615	16.0s
53f368f1	@cf/openai/gpt-oss-20b	5	submitted_passed	yes	Passed	pass	1	16 406	26.8s
d2421f8c	@cf/openai/gpt-oss-20b	6	submitted_failed	yes	Graded	fail	1	10 134	15.3s
b30cd68a	@cf/openai/gpt-oss-20b	7	submitted_passed	yes	Passed	pass	1	9 723	11.8s
1f2c4f23	@cf/openai/gpt-oss-20b	8	submitted_passed	yes	Passed	pass	1	18 720	26.0s
833bcaa3	@cf/openai/gpt-oss-20b	9	submitted_passed	yes	Passed	pass	1	12 440	15.3s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	4	Submitted annotations rejected by the independent grader.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST 811ca1e736f76331d8125a2a 1fd6972080e7f8d7dcdae610 287199470d7896dd	RUN ID live- 20260907T083935Z	STARTED 2026-09- 07T08:39:35.092311 +00:00	FINISHED 2026-09- 07T09:15:48.081959 +00:00
--	--	--	---

Artifact digests

ARTIFACT	SHA-256
publicSummary	7bbfce60747e6238ea44af1f0fd906e699341d60265e8092cbc6eada34840f51
rollouts/0fb1ed25-19d4-434b-accf-9850cc322cd4.json	c27c6e900f2dc55ad3f8818d717a4dafd60bd9fd45e54db7644760d598a041ab
rollouts/1f2c4f23-71e3-4f6d-a2c9-c83665971442.json	cd1cd2afb72f232ce5d87c8f0ea7bfda83d6c0561e689a41de76d8ad3d0bace8
rollouts/21cc70ac-4c9e-4974-8dc7-55777ba722b3.json	826b5e032bd4d4fabe947cf9e911d5676ab913a3ee4054dd7f31da3d2a327833
rollouts/2da20181-72a7-4832-ac72-cb549fc99df9.json	c756231b763c819938ae6bf5e836d621289916344f346482e1204882e7415d6c
rollouts/53f368f1-d8f0-4e27-91b8-f648abda7fc2.json	4b865e8f8757a7a7dd48d402e42ae87ae3e19f9e14dea5d3b121b89dc8f718e2
rollouts/55c7a7d4-6651-4565-b585-75ee1e4d3a51.json	94856a8d6b6409972243ece3d359f3316d9e2b1408001856bc6514db469c9454
rollouts/5756f518-8d8e-497b-8bb3-9d98305faa66.json	4fccc4dbfc72a2af09be8e1ec0b500ab7e779c5067aa463d27b462a09f29b27e
rollouts/6061e717-ae4b-466f-9b21-ded38e70c8c5.json	3da6604ce70b5c68fd42ae32f13c4cf633782b53b9198a0e798a7a678d87c93a
rollouts/6a3337e7-1d52-493e-b53a-8a6f30257d35.json	b2d6934fd4cb55d87b6aa65953bc83fe317aa6281e09fd43c5bffd65b7e7d416
rollouts/6ff0c5ed-b4c7-4e40-806d-bfb4a370c138.json	8315eccdf114d9814b69ce65388547ca8e8a20cd9becffa8bf8af6f5ab46c4cc
rollouts/769d77be-3d37-454d-aa1c-1b65c80e77ff.json	31d48e27de9b230bca3efed430e2c5e7081912c52860fde53e3cd1f570867219
rollouts/77a79aab-2c86-4caa-91e7-975403752155.json	781ca44267427c9091f65fdd2aa5fdf02bd3b01e77edb75d3e09bc1865a40387
rollouts/79a654b1-bdb8-42f9-9c4b-20dd953c5627.json	1cb32b99d66544a11590bdfb010cfb6e7d95c55eb60164d8f81dbf0c8fb7fc2c
rollouts/7b8cbba4-90c3-415e-b94d-6d780b42f8ba.json	eea63d309e3bb2787ff01758ec5e9da267742243338e41bae3ad5fe64811fe57
rollouts/833bcaa3-8d47-48e5-a300-8aaedb7bccfd.json	ef1bee8e4c56c738a4370763c36b47edb58eef1e4a60cf6176476120f664c67b
rollouts/8d664dab-d3e1-43af-b30a-765f9beb9135.json	6ea817cb58f2a1d5614c5c09ac380557fd4755c9fc5afddfb16e03d257501df6
rollouts/8e22d41b-ccf9-4a19-a4b6-16caf07741d1.json	bcff64076d5e017558b0b280f04ba0c45d5c676691fc8cb8874303f6213520be
rollouts/8f0b2bac-c328-4f5c-b28c-4283f200c96d.json	2a0113ab988e1e4d200a68b38f1569f3f6a9b59a06966b4e4e4780990e28ae8b
rollouts/97a444e5-8609-4a78-b30a-23a72f8fa5bb.json	059a7a9eff95ef78f2c457c1715e066dec46f22de8865ab505b78a9c8fd03ab6
rollouts/9f158050-2fba-4365-8f99-ea5b029a1338.json	6828a7a574b7307b74e84a974d45463cdd07dd9241898c9a2be37c7f7335a34d
rollouts/a5066bc6-9497-4d5d-9f5f-732e79d9dcbf.json	b2cda4332f9e664b5575ffc1814d4b4401acacf81c20ba61d9ae20fcf82659c0
rollouts/ae7a3122-fd75-4b14-b8f5-fd069493b36.json	f47c955bc1637b141ffbd7e5e3474ac7cc957647d311332e08fac57544c33b69
rollouts/b2db85b4-95ce-4917-9a9f-05ddf0e47ee4.json	da14f9ce5df3ffff9065b4d7ab13e31fbf680c8e5c329559eb902389416ecfa2
rollouts/b30cd68a-718e-4b9e-b6f6-3763dfccb9d0.json	dc05bb04830799570d1b267279a2cdbf51e25be42b3cbeb422d0f942aa15441b
rollouts/b99939d9-d884-423e-8d6d-e7dce1134325.json	1c80ed0cef763917b05ba02259a47c9ba4aba0a0c3b3385dea160246168ea999
rollouts/d2421f8c-0dfa-41d9-8f54-dbda8a1c5320.json	4d14f7835ad1efcf1953e952b2ec2d996ef2b392ff67fe51fd39ce9e5437da100
rollouts/d8479248-5aa4-41f2-8773-a16295e57890.json	cd83d3e05586c940fabdad52412f3b71209d8e74de39d7547a5ce9840bd09731
rollouts/e650c011-9065-4a93-961b-1e4dab65ba55.json	63ddd545e13195bb0445b1838a2e00336c9bc8433dd401c7ab58a792d7b733a8

ARTIFACT	SHA-256
rollouts/f655f3c1-963f-47d4-8835-605b93026762.json	359c25e87fa0e1edfadcd760d0f8c540cba66ec8540e114062e75ab5b82b6250
rollouts/faa01cab-e8cf-46ae-900d-77237c7d9145.json	ca85be18a75c9c2585eef4f1b7ab759c4c96a23df1d01714ee9c80b1d0ae2e02
spec	f9fe002659962e57f406bfc4f22dd8dfceb5fa49ffe4a302a2fd0d5765f3413d
summary	99c54656aec8bc2f7bba6806063a98f95c32d8abe56dbf209352e046a63d781b

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.

b99939d9 · groq:openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

79a654b1 · groq:openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

769d77be · groq:openai/gpt-oss-20b · 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

STEP	TOOL	PATH	RESULT	CHARS
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.

e650c011 · groq:openai/gpt-oss-20b · 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 (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.

ae7a3122 · groq:openai/gpt-oss-20b · 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.

d8479248 · groq:openai/gpt-oss-20b · seed 5 · submitted_passed

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

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

6a3337e7 · groq:openai/gpt-oss-20b · seed 6 · submitted_passed

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

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

2da20181 · groq:openai/gpt-oss-20b · 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	run_tests	—	ok	423
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	submit	—	ok	85

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

97a444e5 · groq:openai/gpt-oss-20b · seed 8 · submitted_passed

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

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

8e22d41b · groq:openai/gpt-oss-20b · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

f655f3c1 · groq:openai/gpt-oss-120b · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	2 160
3	read_file	annotation-task.json	ok	1 290
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (14 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.

8d664dab · groq:openai/gpt-oss-120b · 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 (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

6061e717 · groq:openai/gpt-oss-120b · 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	2 160
3	read_file	annotation-task.json	ok	1 290
4	edit_file	submission.json	failed	38
5	read_file	submission.json	ok	74
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

6ff0c5ed · groq:openai/gpt-oss-120b · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	2 160
3	read_file	annotation-task.json	ok	1 290

STEP	TOOL	PATH	RESULT	CHARS
4	edit_file	submission.json	failed	38
5	read_file	submission.json	ok	74
6	write_file	submission.json	ok	64
7	run_tests	—	ok	160
8	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.

7b8cbb4 · groq:openai/gpt-oss-120b · seed 4 · submitted_failed

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	run_tests	—	ok	423
4	edit_file	submission.json	failed	38
5	read_file	submission.json	ok	74
6	write_file	submission.json	ok	64
7	run_tests	—	ok	160
8	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.

21cc70ac · groq:openai/gpt-oss-120b · 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	edit_file	submission.json	failed	38
5	read_file	submission.json	ok	74
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	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.

5756f518 · groq:openai/gpt-oss-120b · 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	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (15 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.

a5066bc6 · groq:openai/gpt-oss-120b · 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	run_tests	—	ok	423
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

b2db85b4 · groq:openai/gpt-oss-120b · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (14 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.

0fb1ed25 · groq:openai/gpt-oss-120b · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	280
7	read_file	submission.json	ok	465
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	160
10	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (14 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.

55c7a7d4 · cloudflare:@cf/openai/gpt-oss-20b · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

8f0b2bac · cloudflare:@cf/openai/gpt-oss-20b · seed 1 · submitted_passed

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

STEP	TOOL	PATH	RESULT	CHARS
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	submit	—	ok	85

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

9f158050 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	submit	—	ok	85

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

77a79aab · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	submit	—	ok	85

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

faa01cab · cloudflare:@cf/openai/gpt-oss-20b · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81

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

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

53f368f1 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	invalid_tool_call	(none)	failed	107
6	submit	—	ok	85

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

d2421f8c · cloudflare:@cf/openai/gpt-oss-20b · seed 6 · submitted_failed

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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

b30cd68a · cloudflare:@cf/openai/gpt-oss-20b · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81

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

1f2c4f23 · cloudflare:@cf/openai/gpt-oss-20b · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	write_file	submission.json	ok	64
5	run_tests	—	ok	160
6	invalid_tool_call	(none)	failed	107
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.

833bcaa3 · cloudflare:@cf/openai/gpt-oss-20b · seed 9 · submitted_passed

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

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

10 Attempt record 2 of 2 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260907-aa28bed9 · record digest b5bf2458f67d7fef... · harness clean · containment clean

Executive summary

MODELS TESTED

4

CERTIFIED PASSES

27

of 40 graded rollouts

SUBMITTED

39 of 40

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 40 attempts

Purpose

This report presents a controlled evaluation of 4 model lane(s) against the certified exam vvdex.annotation.structured-data-1 (v0.1.0, family annotation). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

27 of 40

graded rollouts passed the independent hidden tests and were certified

27 of 40 graded rollouts passed. codestral-latest, ministral-14b-latest, ministral-8b-latest, mistral-code-latest 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-01codestral-latest, ministral-14b-latest, ministral-8b-latest, mistral-code-latest passed the independent hidden tests
- F-02submitted_failed × 12 (codestral-latest, ministral-14b-latest, ministral-8b-latest)
- F-03context_exhaustion × 1 (mistral-code-latest)

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: 1944.1s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 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.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	40 (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-aa28bed9
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.structured-data-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:46:48	Evaluation run started
2026-09-07 09:19:12	Evaluation run finished
2026-09-07 09:19:12	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
codestral-latest	mistral	VVDex-configured	VVDex free-quota
mistral-code-latest	mistral	VVDex-configured	VVDex free-quota
ministral-14b-latest	mistral	VVDex-configured	VVDex free-quota
ministral-8b-latest	mistral	VVDex-configured	VVDex free-quota

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 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY http-429	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: 40. Graded (evaluable) rollouts: 40. 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
codestral-latest codestral-latest	10	10	9 / 10	1	0	10 / 10	11.1s	Passed
mistral-code-latest mistral-code-latest	10	10	9 / 10	1	0	9 / 10	35.1s	Passed
ministral-14b-latest ministral-14b-latest	10	10	8 / 10	2	0	10 / 10	43.8s	Passed
ministral-8b-latest ministral-8b-latest	10	10	1 / 10	9	0	10 / 10	25.5s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
codestral-latest	10	10	0	9	90.0%	[60%, 98%]	0–9	90.0%	100.0%	100.0%	100.0%
mistral-code-latest	10	10	0	9	90.0%	[60%, 98%]	0–9	90.0%	100.0%	100.0%	100.0%
ministral-14b-latest	10	10	0	8	80.0%	[49%, 94%]	0–9	80.0%	100.0%	100.0%	100.0%
ministral-8b-latest	10	10	0	1	10.0%	[2%, 40%]	0–9	10.0%	30.0%	50.0%	100.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
codestral-latest	90.0%	[60%, 98%]	101 977	not billed	11.1s	0.0%	submitted_failed
mistral-code-latest	90.0%	[60%, 98%]	307 300	not billed	35.1s	0.0%	context_exhaustion
ministral-14b-latest	80.0%	[49%, 94%]	284 944	not billed	43.8s	0.0%	submitted_failed
ministral-8b-latest	10.0%	[2%, 40%]	193 177	not billed	25.5s	0.0%	submitted_failed

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

This is not a leaderboard. Separated by this run (intervals do not overlap): codestral-latest vs ministral-8b-latest; mistral-code-latest vs ministral-8b-latest; ministral-14b-latest vs ministral-8b-latest. Not separated (intervals overlap): codestral-latest vs mistral-code-latest; codestral-latest vs ministral-14b-latest; mistral-code-latest vs ministral-14b-latest. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	MISTRAL:CODESTRAL-LATEST	MISTRAL:MISTRAL-CODE-LATEST	MISTRAL:MINISTRAL-14B-LATEST	MISTRAL:MINISTRAL-8B-LATEST
Inspect	10 / 10	10 / 10	10 / 10	10 / 10
Annotate	10 / 10	10 / 10	10 / 10	10 / 10
Validate	10 / 10	10 / 10	10 / 10	10 / 10
Submit	10 / 10	9 / 10	10 / 10	10 / 10
Graded	10 / 10	10 / 10	10 / 10	10 / 10
Passed	9 / 10	9 / 10	8 / 10	1 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
mistral:codestral-latest	Passed	Passed	9 / 10	0
mistral:mistral-code-latest	Passed	Passed	9 / 10	0
mistral:ministral-14b-latest	Passed	Passed	8 / 10	0
mistral:ministral-8b-latest	Passed	Graded	1 / 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 28008c20-5f3f-4656-aeb6-e51206eb55ba

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 119ed9c5-c482-41c4-9631-1eed5f893b07

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 4334e358-3059-4c4b-948c-f5c71168ca52

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 28a658b7-f4fd-4318-8f83-7805890e2013

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 7380753e-5293-4688-9e80-3c259dcd2af1

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 3f7f7b06-170d-47c3-8fe5-d230abfcadeb

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 6277d001-0111-492c-b35f-0bef1989c843

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 68efff16-1c75-4ad1-8c84-2dd4300042cc

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 aa78ed1e-d0a3-4d7e-b988-51625bb27fe4

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 a0928ea9-8322-4dfa-9e9f-192c7d75b6c6

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 3406be24-3786-4c88-b898-30f5ffa4467b

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 31c27b11-d445-40e2-9d44-534d51070ba1

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 ced067fc-6392-4d7b-a64a-57fe644d9c0d

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 87cd740f-a5bb-4995-b912-be43e1c73c48

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 28c1c368-43b6-4879-b31e-c6178536a171

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 7005d17c-31a4-4417-9520-095d7677889c

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 fd7054d0-11c6-4002-af9f-4728da35978b

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 fe1efbcb-846f-4861-82e9-d3592f43b597

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 92ad41d1-0aef-43b0-bac7-625a8bc31bd2

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 294a699a-938e-467f-9e1f-22b6e945b8e3

Score	Items	Valid / invalid	Types
0	0	0 / 0	Not recorded

Measured metrics

MetricsNot measured

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

Validation error codes

asset_id, unknown_field

Disagreement

Agreement review

Not measured

Submission 8382639e-0161-42b1-897c-e6ee5ce6268b

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 4dace0db-763f-49b5-9122-5f419ac5e806

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 15969692-86a5-41a3-96e7-cea742326283

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 2722f8a6-5fc7-4459-8b95-a00dc5fa4096

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 f6560efc-c4a7-4806-b8e4-aee938996f7d

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 c5487504-9270-4cf4-88f8-436fca932f79					
Score	Items	Valid / invalid		Types	
0	3	0 / 3		row	
Measured metrics					
Metrics				Not measured	
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
missing_annotation, missing_type, unknown_field					
Disagreement					
Agreement review				Not measured	
Submission 1ab01d1c-d9e8-4380-8a05-8c0feca569a4					
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 cce53f7a-1fa1-424f-a6fc-22a9fef06520

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 8027d7d2-ced8-4ebe-a023-e8e3e43e675f

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 9b6888c6-fa8d-45fa-8e01-8bff3729fb69

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 f77f1667-4e63-40bc-9a18-3e4945e6f378

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 df364322-64a3-4466-bc25-4acb125ef0e6

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 af5feb09-07f7-4f0c-a844-0e909b1c2fd9

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 4ce89da0-a15d-4edf-9b78-808940af2e1c

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 3018d015-bca4-4acf-b5e1-fcd2c64dc1a8

Score	Items	Valid / invalid		Types	
0.75	4	4 / 0		pair, row	

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 62c852d3-f6ad-4140-ab85-70a75ea769a3

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 41a801e0-5807-469a-9397-b18a1fa0d490

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 11daef51-db45-46d5-adb7-6327d0f8514f

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 ce1abde8-711e-4649-82e8-384f3b4ecfdd

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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 95213951-c1c1-44bc-997d-241211e00d3e

Score	Items	Valid / invalid	Types
0.75	4	4 / 0	pair, row

Measured metrics

annotationCount	4	matchedCount	3	mismatchCount	1
pairAccuracy	1	pairF1	1	pairPrecision	1
pairRecall	1	rowAccuracy	0.6667	rowF1	0.6667
rowPrecision	0.6667	rowRecall	0.6667		

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	0.9035	–	–	
rowRecall	recorded	N/A	0.9035	–	–	
rowF1	recorded	N/A	0.9035	–	–	
rowAccuracy	recorded	N/A	0.9035	–	–	

QUALITY SCORE (MEAN)

0.8812

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
mistral:codestral-latest	10	6 / 24	0	0 / 0	0	0	0	0
mistral:mistral-code-latest	10	7 / 24	1	0 / 0	0	0	0	0
mistral:ministral-14b-latest	10	12 / 24	0	0 / 0	0	0	0	0
mistral:ministral-8b-latest	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.

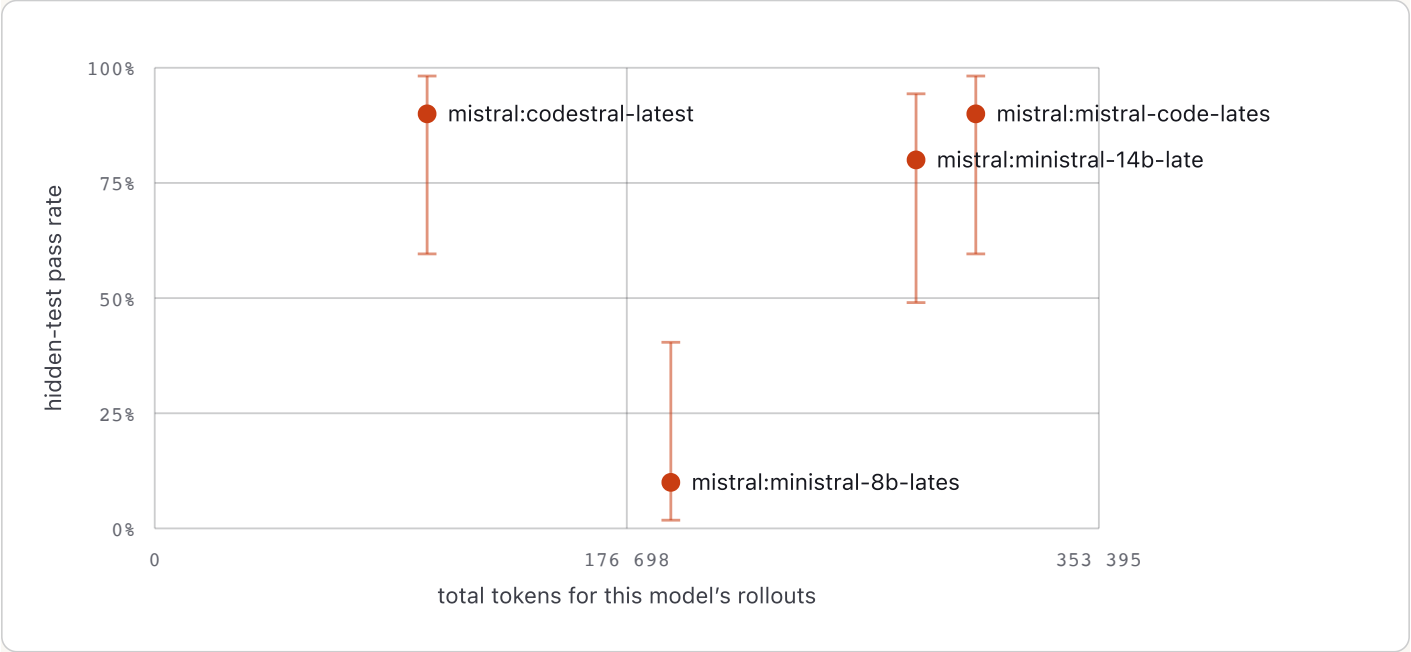
MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
mistral:codestral-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:mistral-code-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:ministral-14b-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:ministral-8b-latest	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 40 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.



Only lanes that report token telemetry are charted. The vertical bar through each point is that model's 95% interval, drawn to the same scale as the axis — where it spans most of the axis, the point is one measurement with a wide interval, not a position on a frontier.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
mistral:codestral-latest	10	101 977	10 197	not billed	9 / 10	90.0%
mistral:mistral-code-latest	10	307 300	30 730	not billed	9 / 10	90.0%
mistral:ministral-14b-latest	10	284 944	28 494	not billed	8 / 10	80.0%
mistral:ministral-8b-latest	10	193 177	19 317	not billed	1 / 10	10.0%

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
codestral-latest, minstral-14b-latest, minstral-8b-latest, mistral-code-latest passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
27 of 40 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 × 12 (codestral-latest, minstral-14b-latest, minstral-8b-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Submitted annotations rejected by the independent grader.	
F-02 · derived from the recorded run data	

F-03	ATTENTION
context_exhaustion × 1 (mistral-code-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Ran out of output context before it could submit.	
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
codestral-latest, ministral-14b-latest, ministral-8b-latest

WHAT WE WOULD LOOK AT FIRST
Review the measured annotation metrics and validator error codes against the declared rubric.

R-02 PROBLEM
Ran out of output context before it could submit.

AFFECTED LANES
mistral-code-latest

WHAT WE WOULD LOOK AT FIRST
The model fills its window before converging — working-set discipline over long files is the gap.

Verification

RUN ID

live-20260907T084648Z

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

b5bf2458f67d7fef214fedfc66497d235bb436dfd5ca1a13080b1eb7f14e00dd

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

SOURCE RECORD FILE

2a98eb3b9200ab6e02a4349affc60d10d4e28d3a70676e9e5eab256b1d2511df

sha256 of `fr-20260907-aa28bed9.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

fe0444ede14b3d22f48b12f704fe1d9c7d350e95a177ed7283b44c104f45036b

sha256 of `fr-20260907-aa28bed9.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

967d96e9819177e802d37613e900dd204d2f5077cf7e9cdd23c4f62df3c09313

sha256 of `fr-20260907-aa28bed9.pdf`.

RUN EVIDENCE BUNDLE

aa28bed9621fe2fdb3687de72d881d68ced68641c5846ef773cc4eeffd5b692

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

CERTIFIED EVIDENCE SEAL

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

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

Measurement history

Previous run on this exam: `fr-20260907-811ca1e7`. 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
28008c20	codestral-latest	0	submitted_failed	yes	Graded	fail	1	12 067	11.7s
119ed9c5	codestral-latest	1	submitted_passed	yes	Passed	pass	1	9 990	10.2s
4334e358	codestral-latest	2	submitted_passed	yes	Passed	pass	1	9 990	10.2s
28a658b7	codestral-latest	3	submitted_passed	yes	Passed	pass	1	9 990	10.3s
7380753e	codestral-latest	4	submitted_passed	yes	Passed	pass	1	9 990	14.5s
3f7f7b06	codestral-latest	5	submitted_passed	yes	Passed	pass	1	9 990	10.7s
6277d001	codestral-latest	6	submitted_passed	yes	Passed	pass	1	9 990	10.3s
68efff16	codestral-latest	7	submitted_passed	yes	Passed	pass	1	9 990	10.1s
aa78ed1e	codestral-latest	8	submitted_passed	yes	Passed	pass	1	9 990	10.6s
a0928ea9	codestral-latest	9	submitted_passed	yes	Passed	pass	1	9 990	11.9s
3406be24	mistral-code-latest	0	submitted_passed	yes	Passed	pass	1	9 879	10.9s
31c27b11	mistral-code-latest	1	submitted_passed	yes	Passed	pass	1	9 990	11.6s
ced067fc	mistral-code-latest	2	submitted_passed	yes	Passed	pass	1	9 990	10.8s
87cd740f	mistral-code-latest	3	submitted_passed	yes	Passed	pass	1	22 062	27.8s
28c1c368	mistral-code-latest	4	submitted_passed	yes	Passed	pass	1	9 990	10.7s
7005d17c	mistral-code-latest	5	submitted_passed	yes	Passed	pass	1	9 990	10.8s
fd7054d0	mistral-code-latest	6	submitted_passed	yes	Passed	pass	1	9 990	9.7s
fe1efbcb	mistral-code-latest	7	submitted_passed	yes	Passed	pass	1	9 990	10.3s
92ad41d1	mistral-code-latest	8	submitted_passed	yes	Passed	pass	1	9 663	8.2s
294a699a	mistral-code-latest	9	context_exhaustion	no	Graded	fail	12	205 756	239.8s
8382639e	ministral-14b-latest	0	submitted_failed	yes	Graded	fail	1	19 369	34.5s
4dace0db	ministral-14b-latest	1	submitted_passed	yes	Passed	pass	1	16 005	31.2s
15969692	ministral-14b-latest	2	submitted_passed	yes	Passed	pass	3	25 711	37.9s
2722f8a6	ministral-14b-latest	3	submitted_passed	yes	Passed	pass	6	43 655	64.5s
f6560efc	ministral-14b-latest	4	submitted_passed	yes	Passed	pass	5	37 741	56.9s
c5487504	ministral-14b-latest	5	submitted_failed	yes	Graded	fail	6	57 671	76.4s
1ab01d1c	ministral-14b-latest	6	submitted_passed	yes	Passed	pass	1	19 174	37.1s
cce53f7a	ministral-14b-latest	7	submitted_passed	yes	Passed	pass	1	16 108	29.5s
8027d7d2	ministral-14b-latest	8	submitted_passed	yes	Passed	pass	4	30 215	44.9s
9b6888c6	ministral-14b-latest	9	submitted_passed	yes	Passed	pass	1	19 295	24.9s
f77f1667	ministral-8b-latest	0	submitted_failed	yes	Graded	fail	2	23 619	26.8s
df364322	ministral-8b-latest	1	submitted_failed	yes	Graded	fail	2	23 676	28.0s
af5feb09	ministral-8b-latest	2	submitted_failed	yes	Graded	fail	2	23 624	30.2s
4ce89da0	ministral-8b-latest	3	submitted_failed	yes	Graded	fail	2	23 630	27.9s
3018d015	ministral-8b-latest	4	submitted_failed	yes	Graded	fail	1	16 259	30.9s

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
62c852d3	ministral-8b-latest	5	submitted_passed	yes	Passed	pass	1	16 287	23.1s
41a801e0	ministral-8b-latest	6	submitted_failed	yes	Graded	fail	1	13 062	18.6s
11daef51	ministral-8b-latest	7	submitted_failed	yes	Graded	fail	2	23 667	27.1s
ce1abde8	ministral-8b-latest	8	submitted_failed	yes	Graded	fail	1	13 089	16.4s
95213951	ministral-8b-latest	9	submitted_failed	yes	Graded	fail	1	16 264	25.5s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	12	Submitted annotations rejected by the independent grader.
context_exhaustion	1	Ran out of output context before it could submit.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST aa28bed9621fe2fdb3687de7 2d881d68cede68641c5846ef 773cc4eefd5b692	RUN ID live- 20260907T084648Z	STARTED 2026-09- 07T08:46:48.76298 9+00:00	FINISHED 2026-09- 07T09:19:12.886352 +00:00
---	--	--	---

Artifact digests

ARTIFACT	SHA-256
publicSummary	952cecc7758a836db4c1897242b5d282d5d3ed3de2b6b6abb1c518d75371c017
rollouts/119ed9c5-c482-41c4-9631-1eed5f893b07.json	a8851a1cc30debd1f09fad72485e5ee2b46cf1b96c7351ab79a84553422edb91
rollouts/11daef51-db45-46d5-adb7-6327d0f8514f.json	3337fd29787a3b3ae3b4e3aacaaf65e531193f1dff847d77759c079cec5e7beb
rollouts/15969692-86a5-41a3-96e7-cea742326283.json	13b4f60c3e6416a33e69097bc8eb02ab4af07954bdbaed7434574fe7c87c04e6
rollouts/1ab01d1c-d9e8-4380-8a05-8c0fec569a4.json	de8afd406f86410cf34f3ac3a45d8444e43fba79e0935fde840b44381f31659e
rollouts/2722f8a6-5fc7-4459-8b95-a00dc5fa4096.json	18bd357a77ab4b8ef05f164abb39efdfbb53ebb7381721fde1752b7d7dab39f2
rollouts/28008c20-5f3f-4656-aeb6-e51206eb55ba.json	b2762e771e33b125d92cbbbbafde4f800b10dfd7924ddc9c522e458c746ce324
rollouts/28a658b7-f4fd-4318-8f83-7805890e2013.json	a50b24e48f8c5e82bfe277ef31ff268eef73e93e194d64329cfae715ec10b6da
rollouts/28c1c368-43b6-4879-b31e-c6178536a171.json	87071fdbe2268ebb7858a935ed029cb6b726c59dc3cc39b5edf893a654849b2d
rollouts/294a699a-938e-467f-9e1f-22b6e945b8e3.json	69861183faa4a6fed844e0244a5a035351188742d33139f18824ea94a40ae4bc
rollouts/3018d015-bca4-4acf-b5e1-fcd2c64dc1a8.json	b14205d5fd9ce38d5efd61cf5065b47e836f77838dbcd9ba670237f01b1124b5
rollouts/31c27b11-d445-40e2-9d44-534d51070ba1.json	e9133714f35da9f73945658d12207b9a2cb9a3fae7f3f5c33a59632d309e6e81
rollouts/3406be24-3786-4c88-b898-30f5ffa4467b.json	20222d9e69def9348c1692662f94f1b2ef6b7153093d5887e64ac242b97fad65
rollouts/3f7f7b06-170d-47c3-8fe5-d230abfcadeb.json	955c9a385216e745c73df98e95f42dbbd0c22d084f2264db7105befd4619b4a5
rollouts/41a801e0-5807-469a-9397-b18a1fa0d490.json	6da0bbe25bf40483f77ed9a578bb7cf4213726e08c63020b7c18f9319f684f0c
rollouts/4334e358-3059-4c4b-948c-f5c71168ca52.json	35b29d0943c6462f8f07bf3013c695001e4ac07b95f0ebe5a7263199b84e3523
rollouts/4ce89da0-a15d-4edf-9b78-808940af2e1c.json	92388939d82fee595e9f05a1010f0b864eec9295f0c48ebb7d9c4943099349d8
rollouts/4dace0db-763f-49b5-9122-5f419ac5e806.json	a14a3ca4c7968f0e838c8134b7ecbcde94265d826188aa7153e97857af204565
rollouts/6277d001-0111-492c-b35f-0bef1989c843.json	ee3bc75f9a8d7877b75ef6fb35ed03929b2159e687caec509fcbdad634b9aff4
rollouts/62c852d3-f6ad-4140-ab85-70a75ea769a3.json	1857b29f68af846648a33a2c380be19a3e2cc966eeba75b3af913de0c5c9df9b
rollouts/68efff16-1c75-4ad1-8c84-2dd4300042cc.json	42e246be8e953e587e60606979cc63d61703af4af101246cebe5e751bf9a2ddf
rollouts/7005d17c-31a4-4417-9520-095d7677889c.json	f7a653c0e972b92d36d028e2ec234bf51e14b01e7393fa96ccdd6383d0626e8
rollouts/7380753e-5293-4688-9e80-3c259dcd2af1.json	2e5ecc54370e1491ce3109a09d02f81b8d41210a34d39b64e3896109f0e3fb49
rollouts/8027d7d2-ced8-4ebe-a023-e8e3e43e675f.json	f865222921a154589572982a763a26ffce5eaf98921f7f9ce48f71be3d0f96e5
rollouts/8382639e-0161-42b1-897c-e6ee5ce6268b.json	a894a15710e94e259f68d383d5251c0d199b558433c8243fd2c9ed8a03bd9b2d
rollouts/87cd740f-a5bb-4995-b912-be43e1c73c48.json	61489fdaff0dd35100357a2be249bdc55c5e34d997239e5fbf6a613f3472f689
rollouts/92ad41d1-0aef-43b0-bac7-625a8bc31bd2.json	b1c2fa033d410b662d4735db06e2e40d90952c44f1d836109dbe31fe8ab29dab
rollouts/95213951-c1c1-44bc-997d-241211e00d3e.json	80ae311962599d56a33e4b02dd72da0bc17b4379f622e8f33486922ef85e74b5
rollouts/9b6888c6-fa8d-45fa-8e01-8bff3729fb69.json	b4ef3904f4c431ebb4783cc70e6ec3fd7302fc7b78ece8031ed5857b0494cd3f

ARTIFACT	SHA-256
rollouts/a0928ea9-8322-4dfa-9e9f-192c7d75b6c6.json	178b710a046bfb0167c5f41b2c4d1da50e72c0fe08537a6f01087559694da0fc
rollouts/aa78ed1e-d0a3-4d7e-b988-51625bb27fe4.json	993c75231e6d8383b454da68bf634d02f11d5a787b43c95412f427152fe5dfe9
rollouts/af5feb09-07f7-4f0c-a844-0e909b1c2fd9.json	926ad982dc8003f2fbdd2763b93364c445023a4f16cc853e803b516d0acab540
rollouts/c5487504-9270-4cf4-88f8-436fca932f79.json	77ececcc3dad0a102edfa00a3dfb5725e5ae3d989edc34dadb11e12fd71c7530
rollouts/cce53f7a-1fa1-424f-a6fc-22a9fef06520.json	775c034d6d27971b59157d21adbb0ddbc1dbcdf9bf4bd7f2bc689c83b3901eb1
rollouts/ce1abde8-711e-4649-82e8-384f3b4ecfdd.json	578e91b322be9bb893df520bdf32b49156a823c7ff4e14110a8eb68ca13d73d2
rollouts/ced067fc-6392-4d7b-a64a-57fe644d9c0d.json	406db65ecad6e3ebd6b76a7d01f45596d872f38f638e5d5ae34c180e2e8b2efd
rollouts/df364322-64a3-4466-bc25-4acb125ef0e6.json	77966184ad035c7f382aa10bc696d277655f4989f6dab3cabb1759bc4f70c70d
rollouts/f6560efc-c4a7-4806-b8e4-aee938996f7d.json	bba069c992524f0be91ca83e1afae667f80cb377951408878cadabc6e4a5d6b4
rollouts/f77f1667-4e63-40bc-9a18-3e4945e6f378.json	e336329bfa242c75ec7f2643eefe84aab049672733ef3bfa97a7fc13a41cc6e4
rollouts/fd7054d0-11c6-4002-af9f-4728da35978b.json	bcc21f7833b4032d72542a33ffbef46f1e9af54ead2b715a89493949c40dd5e0
rollouts/fe1efbcb-846f-4861-82e9-d3592f43b597.json	ac6efcd009cf05fa21b2d48dbbe0df3a2d48b64f46ba1986a331a40fffdcf31a
spec	4d0ddd1ffdd5f68e84ceab928617ac474b2a67a0dd2c6dade020bc1598d9f1c4
summary	53be6404a5b94982d7927eb8fabe97c34954fa80b02248bda5c5a65081c565d0

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.

28008c20 · mistral:codestral-latest · seed 0 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

119ed9c5 · mistral:codestral-latest · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

4334e358 · mistral:codestral-latest · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290

STEP	TOOL	PATH	RESULT	CHARS
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

28a658b7 · mistral:codestral-latest · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

7380753e · mistral:codestral-latest · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3f7f7b06 · mistral:codestral-latest · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160

STEP	TOOL	PATH	RESULT	CHARS
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

6277d001 · mistral:codestral-latest · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

68efff16 · mistral:codestral-latest · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

aa78ed1e · mistral:codestral-latest · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is

the graded answer, so publishing an excerpt of it would publish the answer key.

a0928ea9 · mistral:codestral-latest · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3406be24 · mistral:mistral-code-latest · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

31c27b11 · mistral:mistral-code-latest · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

ced067fc · mistral:mistral-code-latest · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

87cd740f · mistral:mistral-code-latest · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	annotation-task.json	ok	3 290
1	list_files	.	ok	81
2	read_file	submission.json	ok	74
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

28c1c368 · mistral:mistral-code-latest · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

7005d17c · mistral:mistral-code-latest · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

fd7054d0 · mistral:mistral-code-latest · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

fe1efbcb · mistral:mistral-code-latest · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	3 290
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	160
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

92ad41d1 · mistral:mistral-code-latest · seed 8 · submitted_passed

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

294a699a · mistral:mistral-code-latest · seed 9 · context_exhaustion

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	520
1	write_file	submission.json	ok	64
2	run_tests	—	ok	369
3	edit_file	submission.json	ok	53
4	run_tests	—	ok	304
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	304
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	304
9	edit_file	submission.json	ok	53
10	run_tests	—	ok	304
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	304
13	edit_file	submission.json	ok	53
14	run_tests	—	ok	304
15	edit_file	submission.json	ok	53
16	run_tests	—	ok	304
17	edit_file	submission.json	ok	53
18	run_tests	—	ok	304
19	edit_file	submission.json	ok	53
20	run_tests	—	ok	304
21	edit_file	submission.json	ok	53
22	run_tests	—	ok	304
23	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.

8382639e · mistral:ministral-14b-latest · seed 0 · submitted_failed

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	run_tests	—	ok	423
5	run_tests	—	ok	423
6	write_file	submission.json	ok	64
7	run_tests	—	ok	160
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

4dace0db · mistral:ministral-14b-latest · 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	submission.json	ok	74
4	run_tests	—	ok	423
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

15969692 · mistral:ministral-14b-latest · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	assets/records.json	ok	295
3	run_tests	—	ok	423

STEP	TOOL	PATH	RESULT	CHARS
4	edit_file	submission.json	failed	38
5	write_file	submission.json	ok	64
6	run_tests	—	ok	291
7	write_file	submission.json	ok	64
8	run_tests	—	ok	226
9	read_file	annotation-task.json	ok	3 290
10	write_file	submission.json	ok	64
11	run_tests	—	ok	160
12	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

2722f8a6 · mistral:ministral-14b-latest · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	assets/records.json	ok	295
3	write_file	submission.json	ok	64
4	run_tests	—	ok	257
5	write_file	submission.json	ok	64
6	run_tests	—	ok	369
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	304
9	edit_file	submission.json	ok	53
10	run_tests	—	ok	241
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	241
13	read_file	annotation-task.json	ok	3 290
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	160
16	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

f6560efc · mistral:minstral-14b-latest · seed 4 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	assets/records.json	ok	295
3	write_file	submission.json	ok	64
4	run_tests	—	ok	257
5	run_tests	—	ok	257
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	369
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	241
10	edit_file	submission.json	ok	53
11	run_tests	—	ok	241
12	read_file	annotation-task.json	ok	3 290
13	edit_file	submission.json	ok	53
14	run_tests	—	ok	160
15	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

c5487504 · mistral:minstral-14b-latest · seed 5 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	assets/records.json	ok	295
3	run_tests	—	ok	423
4	run_tests	—	ok	423
5	write_file	submission.json	ok	64
6	run_tests	—	ok	354
7	write_file	submission.json	ok	64
8	run_tests	—	ok	283
9	read_file	annotation-task.json	ok	3 290
10	write_file	submission.json	ok	64
11	run_tests	—	ok	390
12	write_file	submission.json	ok	64
13	run_tests	—	ok	390

STEP	TOOL	PATH	RESULT	CHARS
14	write_file	submission.json	ok	64
15	run_tests	—	ok	390
16	run_tests	—	ok	390
17	run_tests	—	ok	390
18	write_file	submission.json	ok	64
19	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (41 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

1ab01d1c · mistral:ministral-14b-latest · 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	submission.json	ok	74
4	run_tests	—	ok	423
5	run_tests	—	ok	423
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

cce53f7a · mistral:ministral-14b-latest · 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	submission.json	ok	74
4	run_tests	—	ok	423
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (37 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

8027d7d2 · mistral:ministral-14b-latest · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	assets/records.json	ok	295
3	write_file	submission.json	ok	64
4	run_tests	—	ok	257
5	run_tests	—	ok	257
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	369
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	304
10	read_file	annotation-task.json	ok	3 290
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	160
13	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

9b6888c6 · mistral:ministral-14b-latest · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	edit_file	submission.json	failed	38
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

f77f1667 · mistral:ministral-8b-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81

STEP	TOOL	PATH	RESULT	CHARS
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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	999
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

df364322 · mistral:minstral-8b-latest · seed 1 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	999
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

af5feb09 · mistral:minstral-8b-latest · seed 2 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	999

STEP	TOOL	PATH	RESULT	CHARS
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

4ce89da0 · mistral:ministral-8b-latest · seed 3 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	999
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160
9	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3018d015 · mistral:ministral-8b-latest · seed 4 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

62c852d3 · mistral:ministral-8b-latest · 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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

41a801e0 · mistral:ministral-8b-latest · seed 6 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

11daef51 · mistral:ministral-8b-latest · seed 7 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	999
7	edit_file	submission.json	ok	53
8	run_tests	—	ok	160

STEP	TOOL	PATH	RESULT	CHARS
9	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

ce1abde8 · mistral:ministral-8b-latest · seed 8 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

95213951 · mistral:ministral-8b-latest · seed 9 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (39 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

11 Attempt record 1 of 2 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-d7a267e3 · record digest 7286bf674cfcf348... · harness clean · containment clean

Executive summary

MODELS TESTED

4

CERTIFIED PASSES

0

of 40 graded rollouts

SUBMITTED

33 of 40

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 40 attempts

Purpose

This report presents a controlled evaluation of 4 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

0 of 40

graded rollouts passed the independent hidden tests and were certified

0 of 40 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 × 33 (codestral-latest, ministral-14b-latest, ministral-8b-latest, mistral-code-latest)
- F-03 invalid_tool_loop × 7 (codestral-latest, mistral-code-latest)

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: 1618.4s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 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.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	40 (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-d7a267e3
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 09:19:13	Evaluation run started
2026-09-07 09:46:12	Evaluation run finished
2026-09-07 09:46:12	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
codestral-latest	mistral	VVDex-configured	VVDex free-quota
mistral-code-latest	mistral	VVDex-configured	VVDex free-quota
ministral-14b-latest	mistral	VVDex-configured	VVDex free-quota
ministral-8b-latest	mistral	VVDex-configured	VVDex free-quota

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 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY http-429	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: 40. Graded (evaluable) rollouts: 40. 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
codestral-latest codestral-latest	10	10	0 / 10	10	0	7 / 10	21.0s	Graded
mistral-code-latest mistral-code-latest	10	10	0 / 10	10	0	6 / 10	22.0s	Graded
ministral-14b-latest ministral-14b-latest	10	10	0 / 10	10	0	10 / 10	22.9s	Graded
ministral-8b-latest ministral-8b-latest	10	10	0 / 10	10	0	10 / 10	17.0s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
codestral-latest	10	10	0	0	0.0%	[0%, 28%]	0–9	0.0%	0.0%	0.0%	0.0%
mistral-code-latest	10	10	0	0	0.0%	[0%, 28%]	0–9	0.0%	0.0%	0.0%	0.0%
ministral-14b-latest	10	10	0	0	0.0%	[0%, 28%]	0–9	0.0%	0.0%	0.0%	0.0%
ministral-8b-latest	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

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
codestral-latest	0.0%	[0%, 28%]	217 876	not billed	21.0s	0.0%	submitted_failed
mistral-code-latest	0.0%	[0%, 28%]	243 696	not billed	22.0s	0.0%	submitted_failed
ministral-14b-latest	0.0%	[0%, 28%]	185 935	not billed	22.9s	0.0%	submitted_failed
ministral-8b-latest	0.0%	[0%, 28%]	128 135	not billed	17.0s	0.0%	submitted_failed

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

This is not a leaderboard. Not separated (intervals overlap): codestral-latest vs mistral-code-latest; codestral-latest vs ministral-14b-latest; codestral-latest vs ministral-8b-latest; mistral-code-latest vs ministral-14b-latest; mistral-code-latest vs ministral-8b-latest; ministral-14b-latest vs ministral-8b-latest. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	MISTRAL:CODESTRAL-LATEST	MISTRAL:MISTRAL-CODE-LATEST	MISTRAL:MINISTRAL-14B-LATEST	MISTRAL:MINISTRAL-8B-LATEST
Inspect	10 / 10	10 / 10	10 / 10	10 / 10
Annotate	10 / 10	10 / 10	10 / 10	10 / 10
Validate	10 / 10	10 / 10	10 / 10	10 / 10
Submit	7 / 10	6 / 10	10 / 10	10 / 10
Graded	10 / 10	10 / 10	10 / 10	10 / 10
Passed	0 / 10	0 / 10	0 / 10	0 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
mistral:codestral-latest	Graded	Graded	0 / 10	0
mistral:mistral-code-latest	Graded	Graded	0 / 10	0
mistral:ministral-14b-latest	Graded	Graded	0 / 10	0
mistral:ministral-8b-latest	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: text. Rubric: 1.0.

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

Submission 4fb67ece-f618-40fc-90cb-f807f6edda26

Score	Items	Valid / invalid	Types
0.7037	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5556	spanPrecision	0.5	spanRecall	0.5

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 46dde59d-7038-40a9-8c30-73021519998c

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 4b3a4997-e625-4169-994c-81c7c48aa711

Score	Items	Valid / invalid	Types
0	0	0 / 0	Not recorded

Measured metrics

Metrics	Not measured
---------	--------------

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

Validation error codes

invalid_submission

Disagreement

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

Submission feede671-eed3-490f-b85a-92ded7aa1659

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 090fd6e1-c44b-4c68-9d54-bd6102534e6a

Score	Items	Valid / invalid	Types
0.7037	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5

spanOverlapIoU	0.5556	spanPrecision	0.5	spanRecall	0.5
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 3601758c-ae6b-401e-890a-dfd8c696676c					
Score	Items	Valid / invalid		Types	
0.75	3	3 / 0		classification, span	
Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5
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 34f7baf1-b120-44a7-9e5c-45c30e1c80cc					
Score	Items	Valid / invalid		Types	
0	0	0 / 0		Not recorded	
Measured metrics					
Metrics				Not measured	
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
invalid_submission					
Disagreement					
Agreement review				Not measured	

Submission 4a293e0d-9686-4397-935f-55a7cb75d856

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 3b259417-ef73-4f01-acb8-498dad01f946

Score	Items	Valid / invalid	Types
0	0	0 / 0	Not recorded

Measured metrics

Metrics	Not measured
---------	--------------

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

Validation error codes

invalid_submission

Disagreement

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

Submission 28d8283a-33c9-4999-bb23-43d932a1925b

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5

spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5
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 16d1434c-af78-41d6-8321-f1ad78315751					
Score	Items	Valid / invalid		Types	
0	0	0 / 0		Not recorded	
Measured metrics					
Metrics					Not measured
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
invalid_submission					
Disagreement					
Agreement review					Not measured
Submission a16db4f7-7a7b-4611-a7f3-83241222d312					
Score	Items	Valid / invalid		Types	
0	0	0 / 0		Not recorded	
Measured metrics					
Metrics					Not measured
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
invalid_submission					
Disagreement					
Agreement review					Not measured
Submission 74bd076b-7c7e-4d7b-9552-92b390ef7bb5					
Score	Items	Valid / invalid		Types	
0.75	3	3 / 0		classification, span	
Measured metrics					

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 2437b75e-39b3-4199-a239-dc53b2bcc9e3

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 3117ef54-8d17-4d7d-8abd-e908000bf940

Score	Items	Valid / invalid	Types
0.8889	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	1

spanOverlapIoU	0.8333	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 4af302af-b01d-4d7b-8020-96ed6ae774c3

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 55ed4113-5dd7-4032-89d5-2fe58129718f

Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	3
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5	spanRecall	0.5

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 5e5adee1-05c5-468e-8e36-7135ddfa5bff			
Score	Items	Valid / invalid	Types
0	0	0 / 0	Not recorded
Measured metrics			
Metrics		Not measured	
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.			
Validation error codes			
invalid_submission			
Disagreement			
Agreement review		Not measured	
Submission af78b2c5-103f-43f7-8995-3796e8302817			
Score	Items	Valid / invalid	Types
0.75	3	3 / 0	classification, span
Measured metrics			
annotationCount	3	classificationAccuracy	1
classificationPrecision	1	classificationRecall	1
mismatchCount	1	spanExactF1	0.5
spanOverlapIoU	0.625	spanPrecision	0.5
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 8636ebd4-5455-4cf7-95bb-54b521b33d2d			
Score	Items	Valid / invalid	Types
0	0	0 / 0	Not recorded

Measured metrics					
Metrics			Not measured		
Track consistency depends on successful box matching. A zero can reflect inaccurate boxes even when track IDs are correct.					
Validation error codes					
invalid_submission					
Disagreement					
Agreement review			Not measured		
Submission 9e02c738-f1a0-4401-8fa2-6dca03be24d0					
Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	
Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5
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 a6768239-65c2-4a3c-a948-e58967a2dc32					
Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	
Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5
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 be56bf5c-45b6-4792-8b7d-0b4e670bf546

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 3464f41e-2621-4a30-b422-9e6d8aa0e1a4

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 91250030-853b-4704-a6d4-664ff8e784b5

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 df3f8d8e-7e6b-4a23-b0bf-a96058870cf5

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 f023a2fe-8944-4cc5-b869-2bb26e03a526

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 9cf78363-721c-4437-88ff-1c2b47dd0642

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 0d2a6c37-bf17-41ed-a550-26a98fbcfe03

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5
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 f1a58bb4-02eb-4a2c-ac8c-8753fed20912					
Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	
Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5
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 62a4ee4a-3939-4f29-9f6e-05fd4616082d					
Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	
Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2

mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 6fbfe8a0-608a-4695-8af0-c5c62cedc58d

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 08812716-b8e5-4030-a026-cc3084208b70

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 9d7ad755-c10d-435e-acf6-1a3dcc9acc40

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 63b96b38-1638-4808-9181-714f29805b89

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 2ece3e4a-a003-4f83-b933-fcdcd9ee97aa

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 5270ce3e-c1dc-4cb9-87ca-3ee4c0f629bc

Score	Items	Valid / invalid		Types	
0.6667	3	3 / 0		classification, span	

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 ae91a5d7-cff3-4bf2-8f78-2c2e8254b3bd

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 a326a97c-2a26-4cce-8dbb-c41b71e52ef7

Score	Items	Valid / invalid	Types
0.6667	3	3 / 0	classification, span

Measured metrics

annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	1	spanExactF1	0.5	spanF1	0.5
spanOverlapIoU	0.5	spanPrecision	0.5	spanRecall	0.5

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 701f1252-1b73-45a8-a393-b4e6328f2784

Score	Items	Valid / invalid	Types
0.6	3	3 / 0	classification, span

Measured metrics					
annotationCount	3	classificationAccuracy	1	classificationF1	1
classificationPrecision	1	classificationRecall	1	matchedCount	2
mismatchCount	2	spanExactF1	0	spanF1	0.5
spanOverlapIoU	0.4	spanPrecision	0.5	spanRecall	0.5

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	0.5152	–	–	
spanRecall	recorded	N/A	0.5152	–	–	
spanF1	recorded	N/A	0.5152	–	–	
spanOverlapIoU	recorded	N/A	0.5445	–	–	
spanExactF1	recorded	N/A	0.4848	–	–	

QUALITY SCORE (MEAN)

0.5745

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
mistral:codestral-latest	10	10 / 24	0	0 / 0	0	0	0	1
mistral:mistral-code-latest	10	11 / 24	0	0 / 0	0	0	0	1
mistral:ministral-14b-latest	10	9 / 24	0	0 / 0	0	0	0	3
mistral:ministral-8b-latest	10	7 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

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

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
mistral:codestral-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:mistral-code-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:ministral-14b-latest	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:ministral-8b-latest	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 40 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.



Only lanes that report token telemetry are charted. The vertical bar through each point is that model's 95% interval, drawn to the same scale as the axis — where it spans most of the axis, the point is one measurement with a wide interval, not a position on a frontier.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
mistral:codestral-latest	10	217 876	21 787	not billed	0 / 10	0.0%
mistral:mistral-code-latest	10	243 696	24 369	not billed	0 / 10	0.0%
mistral:ministral-14b-latest	10	185 935	18 593	not billed	0 / 10	0.0%
mistral:ministral-8b-latest	10	128 135	12 813	not billed	0 / 10	0.0%

Findings

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

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

F-02	ATTENTION
submitted_failed × 33 (codestral-latest, ministral-14b-latest, ministral-8b-latest, mistral-code-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Submitted annotations rejected by the independent grader.	
F-02 · derived from the recorded run data	

F-03	ATTENTION
invalid_tool_loop × 7 (codestral-latest, mistral-code-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Repeatedly called tools that do not exist in this exam.	
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

codestral-latest, minstral-14b-latest, minstral-8b-latest, mistral-code-latest

WHAT WE WOULD LOOK AT FIRST

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

R-02

PROBLEM

Repeatedly called tools that do not exist in this exam.

AFFECTED LANES

codestral-latest, mistral-code-latest

WHAT WE WOULD LOOK AT FIRST

The model cannot hold the stated JSON tool contract (its exact malformed calls are quoted in the transcript appendix) — constrained-format adherence under multi-turn pressure is the gap.

Verification

RUN ID

live-20260907T091913Z

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

7286bf674cfcf3482e06f76daaf1f68a1db1cbf89725833b7c3eb0f9f76e8009

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

SOURCE RECORD FILE

c905cfbc12f420e06416b3701e6eb93c02cca25d6b42280910ccebdccb52b5d9

sha256 of fr-20260907-d7a267e3.json as written beside the standalone report.

SOURCE STANDALONE REPORT

8d9e7d4078e012339814a66c80826638b5dc2554781b343bc3c194d37a6638b2

sha256 of fr-20260907-d7a267e3.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

645cc8b9d813d949c962833b03a9920729a3b87aa483b3704bd8c0073dd9d2aa

sha256 of fr-20260907-d7a267e3.pdf.

RUN EVIDENCE BUNDLE

d7a267e3eb19d0f8894d877fc976fd84f0e92daae9a39b8826819e1fb1180e6a

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

CERTIFIED EVIDENCE SEAL

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

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

Measurement history

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

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

Appendix A — every rollout

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
4fb67ece	codestral-latest	0	submitted_failed	yes	Graded	fail	1	14 147	14.6s
46dde59d	codestral-latest	1	submitted_failed	yes	Graded	fail	1	13 381	13.1s
4b3a4997	codestral-latest	2	invalid_tool_loop	no	Graded	fail	4	42 136	37.5s
feede671	codestral-latest	3	submitted_failed	yes	Graded	fail	1	12 797	14.1s
090fd6e1	codestral-latest	4	submitted_failed	yes	Graded	fail	1	13 555	14.2s
3601758c	codestral-latest	5	submitted_failed	yes	Graded	fail	1	13 627	14.7s
34f7baf1	codestral-latest	6	invalid_tool_loop	no	Graded	fail	4	42 136	38.1s
4a293e0d	codestral-latest	7	submitted_failed	yes	Graded	fail	1	10 580	11.3s
3b259417	codestral-latest	8	invalid_tool_loop	no	Graded	fail	4	42 136	39.2s
28d8283a	codestral-latest	9	submitted_failed	yes	Graded	fail	1	13 381	12.7s
16d1434c	mistral-code-latest	0	invalid_tool_loop	no	Graded	fail	4	42 136	35.3s
a16db4f7	mistral-code-latest	1	invalid_tool_loop	no	Graded	fail	4	42 136	38.5s
74bd076b	mistral-code-latest	2	submitted_failed	yes	Graded	fail	1	13 381	12.7s
2437b75e	mistral-code-latest	3	submitted_failed	yes	Graded	fail	1	13 381	13.4s
3117ef54	mistral-code-latest	4	submitted_failed	yes	Graded	fail	1	11 485	11.6s
4af302af	mistral-code-latest	5	submitted_failed	yes	Graded	fail	1	11 485	11.6s
55ed4113	mistral-code-latest	6	submitted_failed	yes	Graded	fail	1	12 623	12.8s
5e5adee1	mistral-code-latest	7	invalid_tool_loop	no	Graded	fail	4	42 136	34.9s
af78b2c5	mistral-code-latest	8	submitted_failed	yes	Graded	fail	1	13 555	12.9s
8636ebd4	mistral-code-latest	9	invalid_tool_loop	no	Graded	fail	4	41 378	36.0s
9e02c738	ministral-14b-latest	0	submitted_failed	yes	Graded	fail	1	16 232	16.0s
a6768239	ministral-14b-latest	1	submitted_failed	yes	Graded	fail	1	16 232	16.7s
be56bf5c	ministral-14b-latest	2	submitted_failed	yes	Graded	fail	1	16 232	16.0s
3464f41e	ministral-14b-latest	3	submitted_failed	yes	Graded	fail	1	16 232	18.5s
91250030	ministral-14b-latest	4	submitted_failed	yes	Graded	fail	2	29 493	46.0s
df3f8d8e	ministral-14b-latest	5	submitted_failed	yes	Graded	fail	1	26 586	44.8s
f023a2fe	ministral-14b-latest	6	submitted_failed	yes	Graded	fail	1	16 232	16.9s
9cf78363	ministral-14b-latest	7	submitted_failed	yes	Graded	fail	1	16 232	18.0s
0d2a6c37	ministral-14b-latest	8	submitted_failed	yes	Graded	fail	1	16 232	17.4s
f1a58bb4	ministral-14b-latest	9	submitted_failed	yes	Graded	fail	1	16 232	18.5s
62a4ee4a	ministral-8b-latest	0	submitted_failed	yes	Graded	fail	1	11 622	14.1s
6fbfe8a0	ministral-8b-latest	1	submitted_failed	yes	Graded	fail	1	14 451	18.3s
08812716	ministral-8b-latest	2	submitted_failed	yes	Graded	fail	1	14 484	16.7s
9d7ad755	ministral-8b-latest	3	submitted_failed	yes	Graded	fail	1	11 839	15.5s
63b96b38	ministral-8b-latest	4	submitted_failed	yes	Graded	fail	1	11 741	15.0s

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
2ece3e4a	ministral-8b-latest	5	submitted_failed	yes	Graded	fail	1	14 524	21.6s
5270ce3e	ministral-8b-latest	6	submitted_failed	yes	Graded	fail	1	11 746	16.1s
ae91a5d7	ministral-8b-latest	7	submitted_failed	yes	Graded	fail	1	11 671	15.4s
a326a97c	ministral-8b-latest	8	submitted_failed	yes	Graded	fail	1	11 617	16.4s
701f1252	ministral-8b-latest	9	submitted_failed	yes	Graded	fail	1	14 440	20.6s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	33	Submitted annotations rejected by the independent grader.
invalid_tool_loop	7	Repeatedly called tools that do not exist in this exam.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

d7a267e3eb19d0f8894d877f

c976fd84f0e92daae9a39b88

26819e1fb1180e6a

RUN ID

live-

20260907T091913Z

STARTED

2026-09-

07T09:19:13.872782

+00:00

FINISHED

2026-09-

07T09:46:12.274824

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	24838774bc5b699deed1bfbcc3077b1fae780ebc884252207fc2ef45ecac09af
rollouts/08812716-b8e5-4030-a026-cc3084208b70.json	70b86c5e2894bed882746a2c34e38d2554b48b12e3b8860b19baa206525aaaa7
rollouts/090fd6e1-c44b-4c68-9d54-bd6102534e6a.json	229c517846d240026de87a9b502f87ba19a0e8c7ceb6fd4d11bad34a68c80d5b
rollouts/0d2a6c37-bf17-41ed-a550-26a98fbcfef03.json	b7cbfbc13173234bddb4bdc317cd637c87ac99e0d2bf755c8cbf068d8755db4f
rollouts/16d1434c-af78-41d6-8321-f1ad78315751.json	a3975336273253bfdbac5e78aba7ae3dc534cd2bfd42b83cbcf7fca74e39e27d3
rollouts/2437b75e-39b3-4199-a239-dc53b2bcc9e3.json	a04532fe984f812d076267e830c29747325e0fa81cc7184fc14594bac39d79f6
rollouts/28d8283a-33c9-4999-bb23-43d932a1925b.json	066bffff02f823968d7e44ebf82db65fe5161f255f788c3f6ba483cfffdbcd55
rollouts/2ece3e4a-a003-4f83-b933-fcdcd9ee97aa.json	ae668353ed4b2fddc1196c8e74255c7219001e24e3a19ebcde08f788ce788677
rollouts/3117ef54-8d17-4d7d-8abd-e908000bf940.json	2421ccacfac062831a95db8d97212b0d5706d9c26dee91d939221b56658fb3dc
rollouts/3464f41e-2621-4a30-b422-9e6d8aa0e1a4.json	37fc46088cedf68a897694c692212d1a95182b586280034ac42d08ad5e5152ae
rollouts/34f7baf1-b120-44a7-9e5c-45c30e1c80cc.json	46515dfd792773cf84db0432eb4f1776a5caca28a4f5ac2896abcc13c930a8b8
rollouts/3601758c-ae6b-401e-890a-dfd8c696676c.json	dbe97ac7b2fddd45ab8f2093defa1f3bf5a4cd5305448f3314c287afe5160aea
rollouts/3b259417-ef73-4f01-acb8-498dad01f946.json	78fc74082d5fec68d196fc7ceb378b181fe8fcfa96728dac321eaa146d2601a4
rollouts/46dde59d-7038-40a9-8c30-73021519998c.json	9038b73212b57d2259cd851ee24be731bcc4e44e340172b1713be86ad54a0c4b
rollouts/4a293e0d-9686-4397-935f-55a7cb75d856.json	5e29d88280dc959216759da696444b0e073102fd1c9713920f48c818fc8d5ab9
rollouts/4af302af-b01d-4d7b-8020-96ed6ae774c3.json	bf876fc721c98c231c4a50a70caa43abbba28d1d51531716d5d200b055ca8ddd
rollouts/4b3a4997-e625-4169-994c-81c7c48aa711.json	0d8599392c67d303f7cd38e5cbfdc5656830030f2f9964ad0de84c28b729f092
rollouts/4fb67ece-f618-40fc-90cb-f807f6edda26.json	9ac4ed6c803ed66abfc67a9e0aca32f28baa189aca7e75b981ba97c430752c45
rollouts/5270ce3e-c1dc-4cb9-87ca-3ee4c0f629bc.json	0436d6754775d8f6b66b7be2e02692593b6e2c56e488f5a3a1fa369e31821685
rollouts/55ed4113-5dd7-4032-89d5-2fe58129718f.json	e40d87e57dafd8552c23740754f00d7d305a4958795b400664336fa57f48c706
rollouts/5e5adee1-05c5-468e-8e36-7135ddfa5bff.json	f5d37c50ca6dc54e047947837913a8f2631a19588ebe50a693e3149130fff7e43
rollouts/62a4ee4a-3939-4f29-9f6e-05fd4616082d.json	6aeb6b4e43286b42594cf82c5a053e161a51ca772fbdace455bf98becf23a963
rollouts/63b96b38-1638-4808-9181-714f29805b89.json	ab187e93a27f62653ffcc8d993e9fccabd7ea12049d90c42b587ecf0c488b113
rollouts/6fbfe8a0-608a-4695-8af0-c5c62cedc58d.json	d5015fc61520044851b89e513daa33a818d867f39d59e5ef35bc6197a81442e6
rollouts/701f1252-1b73-45a8-a393-b4e6328f2784.json	597d48762d5bd6ab201115412a696cf2b6a63bbadfcdb593819b901a7dcbebc
rollouts/74bd076b-7c7e-4d7b-9552-92b390ef7bb5.json	7ebbeb5292d794b19bed62469bd8f4e1015933f2de74d268c2da48b214cc0594
rollouts/8636ebd4-5455-4cf7-95bb-54b521b33d2d.json	aff2c8c1d1ce5c84f54c59c2f9a7c6062b9765eab0d12587f73642276a76b811
rollouts/91250030-853b-4704-a6d4-664ff8e784b5.json	08a778332899a3dae2f05b3bf33cdceaecdfdc27abdba776f82a31c3896e6f84
rollouts/9cf78363-721c-4437-88ff-1c2b47dd0642.json	ce3783448029f60accee9ae238dfd4fe1646c11950dcfb732cd83174f3c5266b

ARTIFACT	SHA-256
rollouts/9d7ad755-c10d-435e-acf6-1a3dcc9acc40.json	af88b99d1827c763a6e63278be0ac75937f01fdf25ac8d3c068ee46356961a1a
rollouts/9e02c738-f1a0-4401-8fa2-6dca03be24d0.json	2cc0a3c2f9a723ce5b630ac0f5d1b7b5847b729c5819966993335fa177e3cd29
rollouts/a16db4f7-7a7b-4611-a7f3-83241222d312.json	6af442267c94363f6f1d63ce06c991115f0485ae91f31be02968f2a9f15f0311
rollouts/a326a97c-2a26-4cce-8dbb-c41b71e52ef7.json	4afa1f8423a3c5bc6145243c629cd681ef055e6e460548bbc5bd13b3e4c12a85
rollouts/a6768239-65c2-4a3c-a948-e58967a2dc32.json	cc63e4af35cfa156a41456b1a22c0ff0056713631a97ef8d7822296e00432c74
rollouts/ae91a5d7-cff3-4bf2-8f78-2c2e8254b3bd.json	f0c9863ab2c4863e926f2dd703c9e1df896f70bf21c410d6b7c323d02f3b0d09
rollouts/af78b2c5-103f-43f7-8995-3796e8302817.json	4495ad12a7567216c5feb57314d880fb8f37360f9f51698fbf06ce5125001a18
rollouts/be56bf5c-45b6-4792-8b7d-0b4e670bf546.json	7ed84e1e8c029086ddb4f633a3fc14b390101ae246cde8a63bb065cc93800bd
rollouts/df3f8d8e-7e6b-4a23-b0bf-a96058870cf5.json	687a17aac870b8a889345b59873338f8c55393550dbe19b9425af67fe8d89998
rollouts/f023a2fe-8944-4cc5-b869-2bb26e03a526.json	19db2b4e01414c6f7eadbeaa8aaf8955ab3d072a2f63076c85b6e346e3ab6260
rollouts/f1a58bb4-02eb-4a2c-ac8c-8753fed20912.json	31560767f3d114869a249446b98c1fc70c6bbaadd4ef6da41283b0f2b2e15d06
rollouts/feede671-eed3-490f-b85a-92ded7aa1659.json	fecbf6751bff9164ee1d2ba052ac312ce9d3b3403abfa3fd2fcd613aef597f6e
spec	b8e3ae622eeac4cfba46f02b8cb1d54541de6f1c374122859228327bf9a2aa0a
summary	b38dca7573771e785141f03a7237fa754ca230065086e82bf959eef182a81c11

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.

4fb67ece · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	read_file	submission.json	ok	74
2	read_file	annotation-task.json	ok	2 826
3	run_tests	—	ok	442
4	invalid_tool_call	(none)	failed	107
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.

46dde59d · mistral:codestral-latest · seed 1 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	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.

4b3a4997 · mistral:codestral-latest · seed 2 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

feede671 · mistral:codestral-latest · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	assets/message.txt	ok	69
3	read_file	annotation-task.json	ok	2 826
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

090fd6e1 · mistral:codestral-latest · seed 4 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3601758c · mistral:codestral-latest · seed 5 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (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.

34f7baf1 · mistral:codestral-latest · seed 6 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107

STEP	TOOL	PATH	RESULT	CHARS
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

4a293e0d · mistral:codestral-latest · seed 7 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (9 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3b259417 · mistral:codestral-latest · seed 8 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300

STEP	TOOL	PATH	RESULT	CHARS
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

28d8283a · mistral:codestral-latest · seed 9 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	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.

16d1434c · mistral:mistral-code-latest · seed 0 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300
10	invalid_tool_call	(none)	failed	107

STEP	TOOL	PATH	RESULT	CHARS
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

a16db4f7 · mistral:mistral-code-latest · seed 1 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

74bd076b · mistral:mistral-code-latest · seed 2 · submitted_failed

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

STEP	TOOL	PATH	RESULT	CHARS
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

2437b75e · mistral:mistral-code-latest · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	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.

3117ef54 · mistral:mistral-code-latest · seed 4 · submitted_failed

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

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

4af302af · mistral:mistral-code-latest · seed 5 · submitted_failed

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

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

55ed4113 · mistral:mistral-code-latest · seed 6 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	assets/message.txt	ok	69
3	read_file	annotation-task.json	ok	2 826
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (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.

5e5adee1 · mistral:mistral-code-latest · seed 7 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53

STEP	TOOL	PATH	RESULT	CHARS
9	run_tests	—	ok	300
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

af78b2c5 · mistral:mistral-code-latest · seed 8 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

8636ebd4 · mistral:mistral-code-latest · seed 9 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	read_file	assets/message.txt	ok	69
3	read_file	annotation-task.json	ok	2 826
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	300
7	invalid_tool_call	(none)	failed	107
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	300

STEP	TOOL	PATH	RESULT	CHARS
10	invalid_tool_call	(none)	failed	107
11	edit_file	submission.json	ok	53
12	run_tests	—	ok	300
13	invalid_tool_call	(none)	failed	107
14	edit_file	submission.json	ok	53
15	run_tests	—	ok	300
16	invalid_tool_call	(none)	failed	107

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

9e02c738 · mistral:ministral-14b-latest · seed 0 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

a6768239 · mistral:ministral-14b-latest · seed 1 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

be56bf5c · mistral:ministral-14b-latest · seed 2 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

3464f41e · mistral:ministral-14b-latest · seed 3 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

91250030 · mistral:ministral-14b-latest · seed 4 · submitted_failed

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

STEP	TOOL	PATH	RESULT	CHARS
3	read_file	submission.json	ok	74
4	run_tests	—	ok	442
5	run_tests	—	ok	442
6	run_tests	—	ok	442
7	run_tests	—	ok	442
8	edit_file	submission.json	ok	53
9	run_tests	—	ok	316
10	edit_file	submission.json	ok	53
11	run_tests	—	ok	160
12	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

df3f8d8e · mistral:ministral-14b-latest · seed 5 · submitted_failed

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	submission.json	ok	74
4	run_tests	—	ok	442
5	run_tests	—	ok	442
6	run_tests	—	ok	442
7	run_tests	—	ok	442
8	run_tests	—	ok	442
9	edit_file	submission.json	ok	53
10	run_tests	—	ok	160
11	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

f023a2fe · mistral:ministral-14b-latest · seed 6 · submitted_failed

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	submission.json	ok	74

STEP	TOOL	PATH	RESULT	CHARS
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

9cf78363 · mistral:ministral-14b-latest · seed 7 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

0d2a6c37 · mistral:ministral-14b-latest · seed 8 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

f1a58bb4 · mistral:ministral-14b-latest · seed 9 · submitted_failed

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	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	invalid_tool_call	{"tool": "submit"}	failed	119
7	invalid_tool_call	{"tool": "submit"}	failed	119
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

62a4ee4a · mistral:ministral-8b-latest · seed 0 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

6fbfe8a0 · mistral:ministral-8b-latest · seed 1 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

08812716 · mistral:ministral-8b-latest · seed 2 · submitted_failed

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	edit_file	submission.json	failed	38
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

9d7ad755 · mistral:ministral-8b-latest · seed 3 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

63b96b38 · mistral:ministral-8b-latest · seed 4 · submitted_failed

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

STEP	TOOL	PATH	RESULT	CHARS
6	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

2ece3e4a · mistral:ministral-8b-latest · seed 5 · submitted_failed

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	edit_file	submission.json	failed	38
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (32 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.

5270ce3e · mistral:ministral-8b-latest · seed 6 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

ae91a5d7 · mistral:ministral-8b-latest · seed 7 · submitted_failed

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

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

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

a326a97c · mistral:minstral-8b-latest · seed 8 · submitted_failed

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

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

701f1252 · mistral:minstral-8b-latest · seed 9 · submitted_failed

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	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

12 Attempt record 2 of 2 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-dad35536 · record digest d98a068bbbb888bb... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED 3	CERTIFIED PASSES 14 of 14 graded rollouts	SUBMITTED 14 of 14 graded rollouts that called submit	NOT GRADED — LANE ERRORS 12 of 26 attempts
---------------------------------------	--	--	---

Purpose

This report presents a controlled evaluation of 3 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

14 of 14

graded rollouts passed the independent hidden tests and were certified — 12 of 26 attempts not graded (lane errors)

14 of 14 graded rollouts passed. @cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b completed the applicable annotation workflow and were accepted by the independent annotation grader. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 @cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests
- F-02 model_api_failure × 12 (openai/gpt-oss-120b, openai/gpt-oss-20b)
- F-03 The harness suspected itself during this run

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: 5384.2s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 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	26 (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-dad35536
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 09:15:53	Evaluation run started
2026-09-07 10:45:38	Evaluation run finished
2026-09-07 10:45:38	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
openai/gpt-oss-20b	groq	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota
@cf/openai/gpt-oss-20b	cloudflare	VVDex-configured	VVDex free-quota

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 2280s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY http-429	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: 26. Graded (evaluable) rollouts: 14. Not graded — lane errors: 12. 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
openai/gpt-oss-20b openai/gpt-oss-20b	10	1	1 / 1	0	9	1 / 1	61.2s	Passed
openai/gpt-oss-120b openai/gpt-oss-120b	6	3	3 / 3	0	3	3 / 3	73.5s	Passed
@cf/openai/gpt-oss-20b @cf/openai/gpt-oss-20b	10	10	10 / 10	0	0	10 / 10	29.5s	Passed

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
openai/gpt-oss-20b	10	1	9	1	100.0%	[21%, 100%]	0–9	100.0%	—	—	—
openai/gpt-oss-120b	6	3	3	3	100.0%	[44%, 100%]	0–5	100.0%	100.0%	—	—
@cf/openai/gpt-oss-20b	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
openai/gpt-oss-20b	100.0%	[21%, 100%]	9 455	not billed	61.2s	0.0%	model_api_failure
openai/gpt-oss-120b	100.0%	[44%, 100%]	46 261	not billed	73.5s	0.0%	model_api_failure
@cf/openai/gpt-oss-20b	100.0%	[72%, 100%]	126 114	not billed	29.5s	0.0%	none

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

This is not a leaderboard. Not separated (intervals overlap): openai/gpt-oss-20b vs openai/gpt-oss-120b; openai/gpt-oss-20b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-120b vs @cf/openai/gpt-oss-20b. Below the 10-rollout floor for reading a rate as a rate: openai/gpt-oss-20b (1), openai/gpt-oss-120b (3) — read the count, not the percentage. The primary comparison in this report is the stage funnel below — where a model stops is a finding; a percentage over this many rollouts is not.

Stage funnel

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

STAGE	GROQ:OPENAI/GPT-OSS-20B	GROQ:OPENAI/GPT-OSS-120B	CLOUDFLARE:@CF/OPENAI/GPT-OSS-20B
Inspect	1 / 1	3 / 3	10 / 10
Annotate	1 / 1	3 / 3	10 / 10
Validate	1 / 1	3 / 3	10 / 10
Submit	1 / 1	3 / 3	10 / 10
Graded	1 / 1	3 / 3	10 / 10
Passed	1 / 1	3 / 3	10 / 10

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
groq:openai/gpt-oss-20b	Passed	Passed	1 / 1	9
groq:openai/gpt-oss-120b	Passed	Passed	3 / 3	3
cloudflare:@cf/openai/gpt-oss-20b	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 fcdddfc5-451e-4391-acac-d29f767078f3

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 4072a55a-c3d4-4022-a13d-1da0420a80a3

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 286631d0-d089-45db-a6ce-8d8c6c0a8c00

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 1a3bb2c1-b140-4ff5-8252-5b7e10056541

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 ff4cdacd-3a12-42de-9ece-4a33b7afb47

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 c96ed2ed-b42d-4256-82f6-6f4e807a2bbe

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 fca00373-2750-4fd2-9060-7af993f16d86

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 03a6b6b9-d588-46c3-8638-423371f2b7b1

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 ce47ef9f-ca1c-45ae-be0a-95fee49446c4

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 1b23c8d9-356c-49d6-96f1-262508b3e7f6

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 f093a64c-101f-41c3-9128-6696e2430fbc

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 f949bf5d-3025-4002-9223-abf567c5d141

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 1d8d4d41-4f51-40bc-86bd-bc211060a1a6

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 e6c5db54-10ee-4b5b-b5e4-5248e001ae98

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
groq:openai/gpt-oss-20b	10	4 / 24	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	6	6 / 24	0	0 / 0	0	0	0	0
cloudflare:@cf/openai/gpt-oss-20b	10	6 / 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
groq:openai/gpt-oss-20b	10	0	none	0 / 10	0 / 10	0.0%	0
groq:openai/gpt-oss-120b	6	0	none	0 / 6	0 / 6	0.0%	0
cloudflare:@cf/openai/gpt-oss-20b	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 26 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.



Only lanes that report token telemetry are charted. The vertical bar through each point is that model's 95% interval, drawn to the same scale as the axis — where it spans most of the axis, the point is one measurement with a wide interval, not a position on a frontier.

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
groq:openai/gpt-oss-20b	1	9 455	9 455	not billed	1 / 1	100.0%
groq:openai/gpt-oss-120b	3	46 261	15 420	not billed	3 / 3	100.0%
cloudflare:@cf/openai/gpt-oss-20b	10	126 114	12 611	not billed	10 / 10	100.0%

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
@cf/openai/gpt-oss-20b, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
14 of 14 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	LANE / INFRA
model_api_failure × 12 (openai/gpt-oss-120b, openai/gpt-oss-20b)	
Severity: LANE / INFRA	Evidence: Appendix A · Appendix C
The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.	
F-02 · derived from the recorded run data	

F-03	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R1 api-failure share 12/26 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	

Recommendations

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

R-01

PROBLEM

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

AFFECTED LANES

openai/gpt-oss-120b, openai/gpt-oss-20b

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-20260907T091553Z

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

d98a068bbb888bb6215a3e1a72a7d079fc043fd8b98f52d14be0a58303da22e

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

SOURCE RECORD FILE

900b78eed656a05a8ef4faf65d81c10f0e419b2d57bf2a95eb87de4b3c1ab397

sha256 of `fr-20260907-dad35536.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

736960e123f0c51083281c90b6b328fe92c79eb5e9aaa8a32c2b48dd6f21b916

sha256 of `fr-20260907-dad35536.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

8d468df22527c486c36393bbf904f0cb2943a23691aa2996e8ac84287885c464

sha256 of `fr-20260907-dad35536.pdf`.

RUN EVIDENCE BUNDLE

dad3553637a62033fe225b893492f5cca5df0f43d2626c1bbeac643dad71434b

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

CERTIFIED EVIDENCE SEAL

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282

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

Measurement history

Previous run on this exam: `fr-20260907-d7a267e3`. 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
db812d50	openai/gpt-oss-20b	0	model_api_failur e	no	Not graded — lane error	not graded	1	5 688	4.5s
742b65ed	openai/gpt-oss-20b	1	model_api_failur e	no	Not graded — lane error	not graded	1	5 507	32.3s
fccddfc5	openai/gpt-oss-20b	2	submitted_passe d	yes	Passed	pass	1	9 455	61.2s
4caed15c	openai/gpt-oss-20b	3	model_api_failur e	no	Not graded — lane error	not graded	1	7 516	38.9s
3a24ae91	openai/gpt-oss-20b	4	model_api_failur e	no	Not graded — lane error	not graded	1	5 619	42.2s
06fcd034	openai/gpt-oss-20b	5	model_api_failur e	no	Not graded — lane error	not graded	1	5 352	38.7s
e6ba6da1	openai/gpt-oss-20b	6	model_api_failur e	no	Not graded — lane error	not graded	1	5 801	41.7s
2b174252	openai/gpt-oss-20b	7	model_api_failur e	no	Not graded — lane error	not graded	1	5 616	26.0s
8066ac18	openai/gpt-oss-20b	8	model_api_failur e	no	Not graded — lane error	not graded	1	5 791	36.6s
15426cb5	openai/gpt-oss-20b	9	model_api_failur e	no	Not graded — lane error	not graded	1	5 725	896.3s
4072a55a	openai/gpt-oss-120b	0	submitted_passe d	yes	Passed	pass	1	16 282	61.3s
286631d0	openai/gpt-oss-120b	1	submitted_passe d	yes	Passed	pass	1	16 184	86.0s
1a3bb2c1	openai/gpt-oss-120b	2	submitted_passe d	yes	Passed	pass	1	13 795	73.2s
78744a06	openai/gpt-oss-120b	3	model_api_failur e	no	Not graded — lane error	not graded	1	9 800	740.4s
0d8cba8d	openai/gpt-oss-120b	4	model_api_failur e	no	Not graded — lane error	not graded	0	4 568	1513.3s
ef0426f1	openai/gpt-oss-120b	5	model_api_failur e	no	Not graded — lane error	not graded	0	2 680	869.9s
ff4cdacd	@cf/openai/gpt-oss- 20b	0	submitted_passe d	yes	Passed	pass	1	16 576	39.2s
c96ed2ed	@cf/openai/gpt-oss- 20b	1	submitted_passe d	yes	Passed	pass	1	10 854	46.7s
fca00373	@cf/openai/gpt-oss- 20b	2	submitted_passe d	yes	Passed	pass	1	15 977	37.5s
03a6b6b9	@cf/openai/gpt-oss- 20b	3	submitted_passe d	yes	Passed	pass	1	9 768	20.3s
ce47ef9f	@cf/openai/gpt-oss- 20b	4	submitted_passe d	yes	Passed	pass	1	17 513	30.6s
1b23c8d9	@cf/openai/gpt-oss- 20b	5	submitted_passe d	yes	Passed	pass	1	10 300	33.6s

ROLLOUT	MODEL	SEED	OUTCOME	SUBMITTED	DEEPEST STAGE	HIDDEN TESTS	FILES	TOKENS	ELAPSED
f093a64c	@cf/openai/gpt-oss-20b	6	submitted_passed	yes	Passed	pass	1	9 711	16.5s
f949bf5d	@cf/openai/gpt-oss-20b	7	submitted_passed	yes	Passed	pass	1	11 764	30.2s
1d8d4d41	@cf/openai/gpt-oss-20b	8	submitted_passed	yes	Passed	pass	1	11 194	23.6s
e6c5db54	@cf/openai/gpt-oss-20b	9	submitted_passed	yes	Passed	pass	1	12 457	16.5s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	12	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST dad3553637a62033fe225b89 3492f5cca5df0f43d2626c1b beac643dad71434b	RUN ID live- 20260907T091553Z	STARTED 2026-09- 07T09:15:53.797693 +00:00	FINISHED 2026-09- 07T10:45:38.006546 +00:00
--	--	--	---

Artifact digests

ARTIFACT	SHA-256
publicSummary	a730a46832a75775e7be7d4c499934e4de4424569e8658f16d37f6557c470439
rollouts/03a6b6b9-d588-46c3-8638-423371f2b7b1.json	24b9e6ad119952fc4727eeef98a1b5fc9f37b205bd7d2a4527d4a8b2feb1ff2e
rollouts/06fcd034-8a13-4380-8485-c04dd7cfa173.json	f28df6e7f2b4bb47ad310fdfecb63283165accd0699191a46deb99e912998413
rollouts/0d8cba8d-3afd-4eb0-9a7f-c61f3618a072.json	6dfe7bf9fe1f35f069f63dd5d1ad555a8d3035b86bc819060c743f7f2e0fc14e
rollouts/15426cb5-0d21-4eef-9f58-fd54541f6e48.json	ac011dcacacbc27249d5a13970f5d50cdcd3471c35ee6fa0cee36d061ee7b97
rollouts/1a3bb2c1-b140-4ff5-8252-5b7e10056541.json	06ea9115bc7a6f13f05cfd18f2d453db17051379cd6fdb558f1cbb589a95e23d
rollouts/1b23c8d9-356c-49d6-96f1-262508b3e7f6.json	4e48ee344da00ff45597fc4091cc4a81a08230a3c43e205f8e0faee70539a49a
rollouts/1d8d4d41-4f51-40bc-86bd-bc211060a1a6.json	85d5fb7ade4f350162777bc556cc85fa73f99b2446343892639fcbbc104cb6a06
rollouts/286631d0-d089-45db-a6ce-8d8c6c0a8c00.json	25eb5a711a5033d34003f02e496eb28c79fcaa6a2b42bebaa7fba157d4b2da6b
rollouts/2b174252-ac26-4ff7-b84b-777b82aef3d4.json	96e4e7cd4f7fe5947ed35e739614689d833fcbb1689729c6539e21fbfe83e749
rollouts/3a24ae91-a72c-43f3-ae5b-2965d53203cc.json	ee0dbe4ab36a13e16f8c974f1f5d53e2ae5d29dbecc9d9d374b5bf553ceacabd
rollouts/4072a55a-c3d4-4022-a13d-1da0420a80a3.json	b72bdc3be62d6cc3b6815b19f8348fbd1b4b52d4c032ee887a4ebdefb74204f5
rollouts/4caed15c-1741-4d03-9d1f-17883d139e8f.json	12f198ed98f2a26182ca0f5e5bf6e3464c19c8904ad925699c7e1f7368392871
rollouts/742b65ed-7d4c-40eb-83e8-33b2da236d7f.json	7fba9572d20afaaf1d41ca0078c5076634bc3fc0983c8bb96597a8960d04d049
rollouts/78744a06-7809-4afb-a742-9d598907326d.json	aee2f02059c8d9e62bd8d41a1c1d431677a592195081a502d35ce944a70ea8c6
rollouts/8066ac18-a860-4f71-98a3-6632c5b81b87.json	285982303e7ea0e477d6fff781e4f8450e8f4b04a7307e21841b17673d1dc8849
rollouts/c96ed2ed-b42d-4256-82f6-6f4e807a2bbe.json	b1a2fbb8c8955ecf7ab778d3e4011a54e7f35c1b00d9929b86dce84d429b8c06
rollouts/ce47ef9f-ca1c-45ae-be0a-95fee49446c4.json	b4c6dedf368629bcb21b16a72e45e6860c493c4b2eff3c145e97c45026469f33
rollouts/db812d50-90a6-4229-976e-12248ab28fd0.json	fa8b3bc16a0e548ebd108dc387a9de9fe55323e245f2be0c163a268ae09e451e
rollouts/e6ba6da1-46b7-4d65-940c-3085c324410d.json	27f526b476fbc25fc567c87b2cb9e271896e1521bfe03121b59d1e7f3b7fb6e9
rollouts/e6c5db54-10ee-4b5b-b5e4-5248e001ae98.json	e590ae3a09fe598bcfd23cf022d22d5130ba59243aa94486ffb8c33bab88298b
rollouts/ef0426f1-e4ca-44d4-8593-9a544d0de4ee.json	ecf53a0bc257142da75d0f1492881e0d7b7296459960c04e605be4428e53f480
rollouts/f093a64c-101f-41c3-9128-6696e2430fbc.json	928475ecca1f3ce2806af7c57c73f8a73bf762f76276819265988a8c254d3653
rollouts/f949bf5d-3025-4002-9223-abf567c5d141.json	fd140e9baeedf8a60aa04ade62a7383c095a2eae28c4c980d4b00c80588476ba
rollouts/fca00373-2750-4fd2-9060-7af993f16d86.json	a09d8ed709b4984d92b2a1a0f14d369624aca924ccd414188bb7de5ddf2fc36a
rollouts/fccddfc5-451e-4391-acac-d29f767078f3.json	aac49f435861aec8f9493715f411f18c74f7067e60a2c576ee7b7e83d1b7afed
rollouts/ff4cdacd-3a12-42de-9ece-4a33b7afba47.json	90e9b6a069075dede1126442650229de9857738915fff2994031fc5572ab5c47a
spec	049bb22afbd0495e1c5aea226678a2c7993961cbdb5edb38a9cfd42efe5e34ab
summary	9f10f43ec9bc18bb7fe4108c3007100f13759e29e1d63af478b8d39a380f7205

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.

db812d50 · groq:openai/gpt-oss-20b · seed 0 · model_api_failure

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	write_file	submission.json	ok	64

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.

742b65ed · groq:openai/gpt-oss-20b · seed 1 · model_api_failure

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	write_file	submission.json	ok	64

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.

fccddfc5 · groq:openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

4caed15c · groq:openai/gpt-oss-20b · seed 3 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 826
3	read_file	assets/message.txt	ok	69
4	write_file	submission.json	ok	64

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.

3a24ae91 · groq:openai/gpt-oss-20b · seed 4 · model_api_failure

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	write_file	submission.json	ok	64

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.

06fcd034 · groq:openai/gpt-oss-20b · seed 5 · model_api_failure

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	write_file	submission.json	ok	64

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.

e6ba6da1 · groq:openai/gpt-oss-20b · seed 6 · model_api_failure

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	write_file	submission.json	ok	64

Submission diff withheld by the public-disclosure policy (40 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

2b174252 · groq:openai/gpt-oss-20b · seed 7 · model_api_failure

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	write_file	submission.json	ok	64

Submission diff withheld by the public-disclosure policy (16 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.

8066ac18 · groq:openai/gpt-oss-20b · seed 8 · model_api_failure

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	write_file	submission.json	ok	64

Submission diff withheld by the public-disclosure policy (40 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

15426cb5 · groq:openai/gpt-oss-20b · seed 9 · model_api_failure

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	write_file	submission.json	ok	64

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.

4072a55a · groq:openai/gpt-oss-120b · 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	2 160

STEP	TOOL	PATH	RESULT	CHARS
3	read_file	annotation-task.json	ok	826
4	read_file	submission.json	ok	74
5	read_file	assets/message.txt	ok	69
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (29 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

286631d0 · groq:openai/gpt-oss-120b · 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	edit_file	submission.json	failed	38
5	read_file	submission.json	ok	74
6	edit_file	submission.json	ok	53
7	run_tests	—	ok	160
8	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.

1a3bb2c1 · groq:openai/gpt-oss-120b · 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	2 160
3	read_file	annotation-task.json	ok	826
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53
6	run_tests	—	ok	160
7	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (13 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.

78744a06 · groq:openai/gpt-oss-120b · seed 3 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	TASK.md	ok	409
2	read_file	annotation-task.json	ok	2 160
3	read_file	annotation-task.json	ok	826
4	read_file	submission.json	ok	74
5	edit_file	submission.json	ok	53

Submission diff withheld by the public-disclosure policy (13 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.

0d8cba8d · groq:openai/gpt-oss-120b · seed 4 · model_api_failure

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

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

ef0426f1 · groq:openai/gpt-oss-120b · seed 5 · model_api_failure

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

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

ff4cdacd · cloudflare:@cf/openai/gpt-oss-20b · 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	edit_file	submission.json	failed	29
4	read_file	submission.json	ok	74

STEP	TOOL	PATH	RESULT	CHARS
5	edit_file	submission.json	ok	53
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.

c96ed2ed · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

fca00373 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	invalid_tool_call	(none)	failed	107
6	invalid_tool_call	(none)	failed	107
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.

03a6b6b9 · cloudflare:@cf/openai/gpt-oss-20b · 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

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

ce47ef9f · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	invalid_tool_call	(none)	failed	107
5	run_tests	—	ok	160
6	submit	—	ok	85

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

1b23c8d9 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

f093a64c · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64

STEP	TOOL	PATH	RESULT	CHARS
4	run_tests	—	ok	160
5	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.

f949bf5d · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

1d8d4d41 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	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.

e6c5db54 · cloudflare:@cf/openai/gpt-oss-20b · 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	write_file	submission.json	ok	64
4	run_tests	—	ok	160
5	invalid_tool_call	(none)	failed	107

STEP	TOOL	PATH	RESULT	CHARS
6	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (31 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

13 Evidence

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

- `vvdex.annotation.structured-data-1` → `fr-20260907-811ca1e7` · `recordDigest` `50cbe9a820d64927b58510039c580bbbed3bd161565d085363f841eb867181fab`
- `vvdex.annotation.structured-data-1` → `fr-20260907-aa28bed9` · `recordDigest` `b5bf2458f67d7fef214fedfc66497d235bb436dfd5ca1a13080b1eb7f14e00dd`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-d7a267e3` · `recordDigest` `7286bf674cf3482e06f76daaf1f68a1db1cbf89725833b7c3eb0f9f76e8009`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-dad35536` · `recordDigest` `d98a068bbb888bb6215a3e1a72a7d079fc043fd8b98f52d14be0a58303da22e`

14 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.