

Lane pilots — fourteen API lanes and three CLIs on text and structured, one attempt each

2 fixed tasks · 6 sealed attempt records · 25
graded attempts · 13 certified passes

CURRENT

CAMPAIGN REPORT

2026-09-07

Evaluation scope:

cli/codex
cli/claude-sonnet
cli/cursor-grok
mistral-small-latest
mistral-medium-latest
magistral-small-latest
openai/gpt-oss-20b
openai/gpt-oss-120b
codestral-latest
mistral-code-latest
ministral-14b-latest
ministral-8b-latest
@cf/openai/gpt-oss-20b
nvidia/nemotron-3-super-120b-a12b:free
cohere/north-mini-code:free
minimax/minimax-m3:free
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free

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 3 · structured-data-1	12
10.	Attempt record 2 of 3 · structured-data-1	28
11.	Attempt record 3 of 3 · structured-data-1	43
12.	Attempt record 1 of 3 · text-labeling-1	68
13.	Attempt record 2 of 3 · text-labeling-1	92
14.	Attempt record 3 of 3 · text-labeling-1	108
15.	Evidence	123
16.	Attestation	124

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 6	MODEL LANES 17	ATTEMPTS 40
GRADED ATTEMPTS 25	CERTIFIED PASSES 13	LANE ERRORS (NOT GRADED) 15	WITHHELD / INVALID 5

Purpose

This report presents a controlled evaluation of 17 model lane(s) against 2 fixed certified annotation tasks, three 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 three sealed attempt records on each task (6 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

13

certified pass(es) across 25 graded attempts (40 attempts, 15 lane error(s) not graded) in 34 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-0c7e7c67a231 · generated from 6 verified sealed record(s)

Key findings across the campaign

- structured-data-1 · cli/codex · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · cli/claude-sonnet · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · cli/cursor-grok · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (0 lane errors, 0 withheld, 1 invalid).
- structured-data-1 · mistral-small-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · mistral-medium-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · magistral-small-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · openai/gpt-oss-20b · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · openai/gpt-oss-120b · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · codestral-latest · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · mistral-code-latest · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · ministral-14b-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · ministral-8b-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

- structured-data-1 · @cf/openai/gpt-oss-20b · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · nvidia/nemotron-3-super-120b-a12b:free · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (0 lane errors, 1 withheld, 0 invalid).
- structured-data-1 · cohere/north-mini-code:free · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (0 lane errors, 1 withheld, 0 invalid).
- structured-data-1 · minimax/minimax-m3:free · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- structured-data-1 · nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- text-labeling-1 · cli/codex · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/claude-sonnet · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cli/cursor-grok · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (0 lane errors, 0 withheld, 1 invalid).
- text-labeling-1 · mistral-small-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · mistral-medium-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · magistral-small-latest · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 2 excluded (2 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · openai/gpt-oss-20b · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- text-labeling-1 · openai/gpt-oss-120b · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · codestral-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · mistral-code-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · minstral-14b-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · minstral-8b-latest · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · @cf/openai/gpt-oss-20b · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · nvidia/nemotron-3-super-120b-a12b:free · **PASS** — 1 certified passes / 1 graded attempts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · cohere/north-mini-code:free · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (0 lane errors, 1 withheld, 0 invalid).
- text-labeling-1 · minimax/minimax-m3:free · **ATTENTION** — 0 certified passes / 1 graded attempts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- text-labeling-1 · nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free · **NO RESULT** — 0 certified passes / 0 graded attempts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	ANNOTATION · STRUCTURED-DATA-1 VVDEX.ANNOTATION.STRUCTURED-DAT A-1	ANNOTATION · TEXT-LABELING-1 VVDEX.ANNOTATION.TEXT-LABELIN G-1
@cf/openai/gpt-oss-20b	1/1	1/1
cli/claude-sonnet	1/1	1/1
cli/codex	1/1	1/1
openai/gpt-oss-120b	1/1	1/1
codestral-latest	1/1	0/1
minimax/minimax-m3:free	1/1	0/1
mistral-code-latest	1/1	0/1
nvidia/nemotron-3-super-120b-a12b:free	0/0 + 1 withheld	1/1
openai/gpt-oss-20b	1/1	0/0 + 1 lane error
cli/cursor-grok	0/0 + 1 invalid	0/0 + 1 invalid
cohere/north-mini-code:free	0/0 + 1 withheld	0/0 + 1 withheld
magistral-small-latest	0/0 + 2 lane errors	0/0 + 2 lane errors
ministral-14b-latest	0/1	0/1
ministral-8b-latest	0/1	0/1
mistral-medium-latest	0/0 + 2 lane errors	0/0 + 2 lane errors
mistral-small-latest	0/0 + 2 lane errors	0/0 + 2 lane errors
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0/0 + 1 lane error	0/0 + 1 lane error

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

03 Environment families

This campaign covers 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	2	2	0	0	0
cli/claude-sonnet	2	2	0	0	0
cli/codex	2	2	0	0	0
openai/gpt-oss-120b	2	2	0	0	0
codestral-latest	1	2	0	0	0
minimax/minimax-m3:free	1	2	0	0	0
mistral-code-latest	1	2	0	0	0
nvidia/nemotron-3-super-120b-a12b:free	1	1	0	1	0
openai/gpt-oss-20b	1	1	1	0	0
cli/cursor-grok	0	0	0	0	2
cohere/north-mini-code:free	0	0	0	2	0
magistral-small-latest	0	0	4	0	0
ministral-14b-latest	0	2	0	0	0
ministral-8b-latest	0	2	0	0	0
mistral-medium-latest	0	0	4	0	0
mistral-small-latest	0	0	4	0	0
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	0	2	0	0

Raw counts first. *Graded attempts* = certified passes + model fails; lane errors, withheld and invalid cells are outside capability arithmetic in both directions. These 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

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

structured-data-1 · R4 · row-level

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

structured-data-1 · R1 · diagnostic

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

text-labeling-1 · R1 · diagnostic

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

text-labeling-1 · R4 · row-level

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

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
structured-data-1	cli/cursor-grok	0	INVALID	isolation_breach
structured-data-1	cohere/north-mini-code:free	0	withheld	withheld (conditions unproven)
structured-data-1	magistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	magistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	mistral-medium-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	mistral-medium-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	mistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	mistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
structured-data-1	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	lane error	HTTP 429: Rate limit exceeded: free-models-per-day. Add 5 credits to unlock 1000 free model requests per day
structured-data-1	nvidia/nemotron-3-super-120b-a12b:free	0	withheld	withheld (conditions unproven)
text-labeling-1	cli/cursor-grok	0	INVALID	isolation_breach
text-labeling-1	cohere/north-mini-code:free	0	withheld	withheld (conditions unproven)
text-labeling-1	magistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	magistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	mistral-medium-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	mistral-medium-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	mistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	mistral-small-latest	0	lane error	HTTP 429: Rate limit exceeded
text-labeling-1	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	lane error	HTTP 429: Rate limit exceeded: free-models-per-day. Add 5 credits to unlock 1000 free model requests per day
text-labeling-1	openai/gpt-oss-20b	0	lane error	HTTP 400: Failed to parse tool call arguments as JSON

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 3 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.
text-labeling-1	answer_key	Applied consistently to 3 sealed records. answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family annotation): the exam is VVDex evaluation IP and the graded output is an answer or a keyed value, so no submission content, diff, model claim or hidden-assertion name is published.

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

vvdex.annotation.structured-data-1 · record fr-20260907-15d51112 · record digest d074d0dd0c2e7a85... · harness SUSPECT · {"breached": 1, "unverified": 0}

Executive summary

MODELS TESTED

3

CERTIFIED PASSES

2

of 3 graded rollouts

SUBMITTED

3 of 3

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 3 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

2 of 3

graded rollouts passed the independent hidden tests and were certified

2 of 3 graded rollouts passed. cli/claude-sonnet, cli/codex completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 3 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was isolation_breach — .

Key findings

- F-01 cli/claude-sonnet, cli/codex passed the independent hidden tests
- F-02 isolation_breach × 1 (cli/cursor-grok)
- F-03 1 rollout(s) acted outside the exam channel — isolation breach
- F-04 The harness suspected itself during this run
- F-05 Statistical floor — 3 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 155.0s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 14520s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

The instrument. Exam `vvdex.annotation.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	3 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

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

Evaluation configuration

REPORT	fr-20260907-15d51112
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 08:47:44	Evaluation run started
2026-09-07 08:50:19	Evaluation run finished
2026-09-07 08:50:19	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Cursor Agent CLI (Grok 4.6 Fast, subscription)	cli	local runtime	operator subscription, flat — not billed per token

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

Limits

RUNTIME docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 3. Graded (evaluable) rollouts: 3. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Codex CLI (GPT, subscription) cli/codex	1	1	1 / 1	0	0	1 / 1	62.5s	Passed
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	1	1	1 / 1	0	0	1 / 1	19.4s	Passed
Cursor Agent CLI (Grok 4.6 Fast, subscription) cli/cursor-grok	1	1	0 / 1	1	0	1 / 1	71.1s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
cli/codex	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/cursor-grok	1	1	0	0	0.0%	[0%, 79%]	0	0.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/codex	100.0%	[21%, 100%]	n/a	n/a	62.5s	0.0%	none
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	19.4s	0.0%	none
cli/cursor-grok	0.0%	[0%, 79%]	n/a	n/a	71.1s	0.0%	isolation_breach

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

This is not a leaderboard. Not separated (intervals overlap): cli/codex vs cli/claude-sonnet; cli/codex vs cli/cursor-grok; cli/claude-sonnet vs cli/cursor-grok. Below the 10-rollout floor for reading a rate as a rate: cli/codex (1), cli/claude-sonnet (1), cli/cursor-grok (1) — 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	CLI:CLI/CODEX	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CURSOR-GROK
Inspect	1 / 1	1 / 1	1 / 1
Annotate	1 / 1	1 / 1	1 / 1
Validate	1 / 1	1 / 1	1 / 1
Submit	1 / 1	1 / 1	1 / 1
Graded	1 / 1	1 / 1	1 / 1
Passed	1 / 1	1 / 1	0 / 1

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/codex	Passed	Passed	1 / 1	0
cli:cli/claude-sonnet	Passed	Passed	1 / 1	0
cli:cli/cursor-grok	Graded	Graded	0 / 1	0

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

Annotation evidence

Modality: structured. **Rubric:** 1.0.

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

Submission 2c8b8df7-cb4f-48d8-820b-e111f13e1472

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 2ed89e7a-80cf-458e-8163-8ed6837cab40

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 91496ee2-7292-4b0a-b13a-a28e2f701999

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

Measured metrics

annotationCount	4	matchedCount	4	mismatchCount	0
pairAccuracy	1	pairF1	1	pairPrecision	1

pairRecall	1	rowAccuracy	1	rowF1	1
rowPrecision	1	rowRecall	1		

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

Validation error codes

None recorded

Disagreement

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

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

Quality axes

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

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
pairPrecision	recorded	N/A	1	–	–	
pairRecall	recorded	N/A	1	–	–	
pairF1	recorded	N/A	1	–	–	
pairAccuracy	recorded	N/A	1	–	–	
rowPrecision	recorded	N/A	1	–	–	
rowRecall	recorded	N/A	1	–	–	
rowF1	recorded	N/A	1	–	–	
rowAccuracy	recorded	N/A	1	–	–	

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
cli:cli/codex	1	8 / 24	0	0 / 0	0	0	0	0
cli:cli/claude-sonnet	1	5 / 24	0	0 / 0	0	0	0	0
cli:cli/cursor-grok	1	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
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/cursor-grok	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 3 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	PASS
cli/claude-sonnet, cli/codex passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
2 of 3 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
isolation_breach × 1 (cli/cursor-grok)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
No description recorded for this class.	
F-02 · derived from the recorded run data	

F-03	ATTENTION
1 rollout(s) acted outside the exam channel — isolation breach	
Severity: ATTENTION	Evidence: Appendix A
The taker used its own tools, memory or network rather than the exam's actions. Affected lane(s): cli/cursor-grok. 1 pass(es) were withdrawn: a grader exit code earned outside the channel is not a model result, and this report does not present it as one.	
F-03 · derived from the recorded run data	

F-04	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R4 containment breach in 1/3 rollouts (lanes ['cli/cursor-grok']) — takers acted outside the exam channel; those rows are not model results. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-04 · derived from the recorded run data	

F-05	NOTE
Statistical floor — 3 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-05 · derived from the recorded run data	

Recommendations

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

R-01 For anyone reading the percentages

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

Verification

RUN ID

live-20260907T084744Z

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

d074d0dd0c2e7a855207c95b302bfc b08bc7a092265803601f8a19c28ea6a2bb

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

SOURCE RECORD FILE

a05c7d157b4c824fd4d8c38b1fd0be4d2b3916eb9cad877bfc1323be6062131c

sha256 of `fr-20260907-15d51112.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

7725cc4fee98ec1e91403e13ca9e1ff5600b59f68ff993d65f2275f33945f717

sha256 of `fr-20260907-15d51112.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

46d3d53d519f5d6bb4664114ffa4d7e6405549d06b14072aabfded6a979459bd

sha256 of `fr-20260907-15d51112.pdf`.

RUN EVIDENCE BUNDLE

15d51112110a2259a6a4dc53cdcfc35fa9b73dafc0cfd81e97ed997391ab6b929

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-20260907-29701401`. 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
2c8b8df7	cli/codex	0	submitted_passed	yes	Passed	pass	1	n/a	62.5s
2ed89e7a	cli/claude-sonnet	0	submitted_passed	yes	Passed	pass	1	n/a	19.4s
91496ee2	cli/cursor-grok	0	submitted_passed	yes	Graded	fail	1	n/a	71.1s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

15d51112110a2259a6a4dc53cd

cf35fa9b73dafc0cfd81e97ed9

97391ab6b929

RUN ID

live-

20260907T084744Z

STARTED

2026-09-

07T08:47:44.828836+

00:00

FINISHED

2026-09-

07T08:50:19.847044+

00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	c1b46a0f20d006929db0bcea8732c1487781f7df0b01c66e16cb883b83131559
rollouts/2c8b8df7-cb4f-48d8-820b-e111f13e1472.json	659ffd18ae4beafb2810273c24dff49e6065984b28d90185f220bc43f442af63
rollouts/2ed89e7a-80cf-458e-8163-8ed6837cab40.json	dcbfc3ea963bd74cdda125f15973cf7f0e562a768e7a00f54a4b8d633030db8c
rollouts/91496ee2-7292-4b0a-b13a-a28e2f701999.json	3a5a549a0fe1212d98139b8730eac2680321e7b9ec0c1367f2567d10d98b0009
spec	67000daeab2d91eff6c10be6c9f7bd8297501c3db4faa856069becb29f6f0da5
summary	b7d1e67f73015684eb5a19d71edbd448f39c61558ecbc6335e0779869b178fb

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.

2c8b8df7 · cli:cli/codex · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	read_file	TASK.md	ok	520
2	read_file	annotation-task.json	ok	3 290
3	read_file	assets/records.json	ok	295
4	read_file	submission.json	ok	74
5	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.

2ed89e7a · cli:cli/claude-sonnet · seed 0 · submitted_passed

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

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

91496ee2 · cli:cli/cursor-grok · 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	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 3 · structured-data-1

vvdex.annotation.structured-data-1 · record fr-20260907-29701401 · record digest 4d674e21c5842739... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

3

CERTIFIED PASSES

0

of 0 graded rollouts

SUBMITTED

0 of 0

graded rollouts that called submit

NOT GRADED — LANE ERRORS

3

of 3 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

0 of 0

graded rollouts passed the independent hidden tests and were certified — 3 of 3 attempts not graded (lane errors)

0 of 0 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 3 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 model_api_failure × 3 (magistral-small-latest, mistral-medium-latest, mistral-small-latest)
- F-03 The harness suspected itself during this run
- F-04 Statistical floor — 3 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 216.8s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.structured-data-1	0.1.0	annotation	3 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

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

Evaluation configuration

REPORT	fr-20260907-29701401
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 08:43:11	Evaluation run finished
2026-09-07 08:43:11	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
mistral-small-latest	mistral	VVDex-configured	VVDex free-quota
mistral-medium-latest	mistral	VVDex-configured	VVDex free-quota
magistral-small-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: 3. Graded (evaluable) rollouts: 0. Not graded — lane errors: 3. 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
mistral-small-latest mistral-small-latest	1	0	not graded	0	1	—	—	nothing
mistral-medium-latest mistral-medium-latest	1	0	not graded	0	1	—	—	nothing
magistral-small-latest magistral-small-latest	1	0	not graded	0	1	—	—	nothing

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS
mistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0
mistral-medium-latest	1	0	1	0	not recorded	[0%, 100%]	0
magistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
mistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
mistral-medium-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
magistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure

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): mistral-small-latest vs mistral-medium-latest; mistral-small-latest vs magistral-small-latest; mistral-medium-latest vs magistral-small-latest. Below the 10-rollout floor for reading a rate as a rate: mistral-small-latest (0), mistral-medium-latest (0), magistral-small-latest (0) — 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	MISTRAL:MISTRAL-SMALL-LATEST	MISTRAL:MISTRAL-MEDIUM-LATEST	MISTRAL:MAGISTRAL-SMALL-LATEST
Inspect	0 / 0	0 / 0	0 / 0
Annotate	0 / 0	0 / 0	0 / 0
Validate	0 / 0	0 / 0	0 / 0
Submit	0 / 0	0 / 0	0 / 0
Graded	0 / 0	0 / 0	0 / 0
Passed	0 / 0	0 / 0	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
mistral:mistral-small-latest	nothing	nothing	0 / 0	1
mistral:mistral-medium-latest	nothing	nothing	0 / 0	1
mistral:magistral-small-latest	nothing	nothing	0 / 0	1

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

Annotation evidence

Modality: structured. **Rubric:** 1.0.

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

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

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
mistral:mistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:mistral-medium-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:magistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

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

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
mistral:mistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-medium-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:magistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 3 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

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

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
mistral:mistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:mistral-medium-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:magistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

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

F-02	LANE / INFRA
model_api_failure × 3 (magistral-small-latest, mistral-medium-latest, mistral-small-latest)	
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 3/3 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	

F-04	NOTE
Statistical floor — 3 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-04 · derived from the recorded run data	

Recommendations

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

R-01 PROBLEM

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

AFFECTED LANES

magistral-small-latest, mistral-medium-latest, mistral-small-latest

WHAT WE WOULD LOOK AT FIRST

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

R-02 For anyone reading the percentages

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

Verification

RUN ID

live-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

4d674e21c58427397f8b9b5918e587f9111427e72eaa2281bb7d0fcfa7bd7848

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

SOURCE RECORD FILE

9368ab53d70567c8057070fb70123ddd9624bf3fb0a2adb593f804da2995c9e5

sha256 of `fr-20260907-29701401.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

ddc519f68ae03a1ef094f4947ca3c4bd31c79b58dfa43585a3efbc65e997de9d

sha256 of `fr-20260907-29701401.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

08cdc6a768da4fb5eccc38a2ff1bc326e503dd20fc2f338ba72bb9446c8579d5

sha256 of `fr-20260907-29701401.pdf`.

RUN EVIDENCE BUNDLE

29701401b233d331cd37410b9dbf5d57ec5b9b5b7e235e2b30b64e4a4a119765

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-20260907-7c027ac8`. 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
5325ee0d	mistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.2s
ba4b5d6d	mistral-medium-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.3s
348b5a59	magistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	59.3s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	3	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST
29701401b233d331cd37410b9d
bf5d57ec5b9b5b7e235e2b30b6
4e4a4a119765

RUN ID
live-
20260907T083935Z

STARTED
2026-09-
07T08:39:35.083697+
00:00

FINISHED
2026-09-
07T08:43:11.912785+
00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	55ac1823a1f4949fe9a3f50b163f57c5e213360ea76be2609e3cafaece9db4f6
rollouts/348b5a59-2461-4db4-bda8-14986f536fed.json	e3c4ba6d67cca0b40c3ca0771ddac19675cff86c44c1d0671ec303c9235ef8ca
rollouts/5325ee0d-5a54-44a7-bd13-f93292b6e6b3.json	76847c3654604565e09e5823b3ebe7603105f9f679c473b3584d5778a1dc373b
rollouts/ba4b5d6d-3b72-42ef-b268-e03d0b28d4be.json	9254e4d7f3e1fd79be271c8a69c9330647462c9c961d3bf24d288c21e224adcf
spec	cc52469b6ba5dcb93db6b639c91bbb8021e551b167568697411aa2983f6fcd39
summary	77fe6c15226e01f4dbaf612bea4618e9a54e2956837b72f7cc273f01b3eac732

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.

5325ee0d · mistral:mistral-small-latest · seed 0 · model_api_failure

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

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

ba4b5d6d · mistral:mistral-medium-latest · seed 0 · model_api_failure

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

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

348b5a59 · mistral:magistral-small-latest · seed 0 · model_api_failure

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

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

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

vvdex.annotation.structured-data-1 · record fr-20260907-7c027ac8 · record digest 4eb1c3548b8b094d... · harness clean · containment clean

Executive summary

MODELS TESTED

14

CERTIFIED PASSES

6

of 10 graded rollouts

SUBMITTED

8 of 10

graded rollouts that called submit

NOT GRADED — LANE ERRORS

4

of 14 attempts

Purpose

This report presents a controlled evaluation of 14 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

6 of 10

graded rollouts passed the independent hidden tests and were certified — 4 of 14 attempts not graded (lane errors)

6 of 10 graded rollouts passed. @cf/openai/gpt-oss-20b, codestral-latest, minimax/minimax-m3:free, mistral-code-latest, 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, codestral-latest, minimax/minimax-m3:free, mistral-code-latest, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests
- F-02 model_api_failure × 4 (magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free)
- F-03 submitted_failed × 2 (ministral-14b-latest, ministral-8b-latest)
- F-04 incomplete_submission × 1 (nvidia/nemotron-3-super-120b-a12b:free)
- F-05 invalid_tool_loop × 1 (cohere/north-mini-code:free)

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: 987.0s.

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

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.structured-data-1	0.1.0	annotation	14 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

6844331643ab3301e2c3cb129edef1ca2bc06b5d0baaa69ba0b3762c6a941c4c

EXAM FINGERPRINT

5d6b3708dba2cdf08218a449...

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

Evaluation configuration

REPORT	fr-20260907-7c027ac8
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 08:21:38	Evaluation run started
2026-09-07 08:38:05	Evaluation run finished
2026-09-07 08:38:05	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
mistral-small-latest	mistral	VVDex-configured	VVDex free-quota
codestral-latest	mistral	VVDex-configured	VVDex free-quota
mistral-code-latest	mistral	VVDex-configured	VVDex free-quota
mistral-medium-latest	mistral	VVDex-configured	VVDex free-quota
magistral-small-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
@cf/openai/gpt-oss-20b	cloudflare	VVDex-configured	VVDex free-quota
nvidia/nemotron-3-super-120b-a12b:free	openrouter	VVDex-configured	VVDex free-quota
cohere/north-mini-code:free	openrouter	VVDex-configured	VVDex free-quota
minimax/minimax-m3:free	openrouter	VVDex-configured	VVDex free-quota
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	openrouter	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: 14. Graded (evaluable) rollouts: 10. Not graded — lane errors: 4. 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	1	1	1 / 1	0	0	1 / 1	74.2s	Passed
openai/gpt-oss-120b openai/gpt-oss-120b	1	1	1 / 1	0	0	1 / 1	134.8s	Passed
mistral-small-latest mistral-small-latest	1	0	not graded	0	1	—	—	nothing
codestral-latest codestral-latest	1	1	1 / 1	0	0	1 / 1	10.0s	Passed
mistral-code-latest mistral-code-latest	1	1	1 / 1	0	0	1 / 1	9.5s	Passed
mistral-medium-latest mistral-medium-latest	1	0	not graded	0	1	—	—	nothing
magistral-small-latest magistral-small-latest	1	0	not graded	0	1	—	—	nothing
ministral-14b-latest ministral-14b-latest	1	1	0 / 1	1	0	1 / 1	78.7s	Graded
ministral-8b-latest ministral-8b-latest	1	1	0 / 1	1	0	1 / 1	20.9s	Graded
@cf/openai/gpt-oss-20b @cf/openai/gpt-oss-20b	1	1	1 / 1	0	0	1 / 1	27.7s	Passed
nvidia/nemotron-3-super-120b-a12b:free nvidia/nemotron-3-super-120b-a12b:free	1	1	0 / 1	1	0	0 / 1	111.8s	Graded
cohere/north-mini-code:free cohere/north-mini-code:free	1	1	0 / 1	1	0	0 / 1	29.7s	Graded
minimax/minimax-m3:free minimax/minimax-m3:free	1	1	1 / 1	0	0	1 / 1	38.5s	Passed
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	not graded	0	1	—	—	nothing

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
openai/gpt-oss-20b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
openai/gpt-oss-120b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
mistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
codestral-latest	1	1	0	1	100.0%	[21%, 100%]	0	100.0%

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
mistral-code-latest	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
mistral-medium-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
magistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
ministral-14b-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
ministral-8b-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
@cf/openai/gpt-oss-20b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
nvidia/nemotron-3-super-120b-a12b:free	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
cohere/north-mini-code:free	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
minimax/minimax-m3:free	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	1	0	not recorded	[0%, 100%]	0	—

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; 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%]	9 985	not billed	74.2s	0.0%	none
openai/gpt-oss-120b	100.0%	[21%, 100%]	17 309	not billed	134.8s	0.0%	none
mistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
codestral-latest	100.0%	[21%, 100%]	9 990	not billed	10.0s	0.0%	none
mistral-code-latest	100.0%	[21%, 100%]	9 879	not billed	9.5s	0.0%	none
mistral-medium-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
magistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
ministral-14b-latest	0.0%	[0%, 79%]	69 188	not billed	78.7s	0.0%	submitted_failed
ministral-8b-latest	0.0%	[0%, 79%]	16 227	not billed	20.9s	0.0%	submitted_failed
@cf/openai/gpt-oss-20b	100.0%	[21%, 100%]	18 668	not billed	27.7s	0.0%	none
nvidia/nemotron-3-super-120b-a12b:free	0.0%	[0%, 79%]	11 277	not billed	111.8s	0.0%	incomplete_submission
cohere/north-mini-code:free	0.0%	[0%, 79%]	7 756	not billed	29.7s	0.0%	invalid_tool_loop
minimax/minimax-m3:free	100.0%	[21%, 100%]	17 386	not billed	38.5s	0.0%	none
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure

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 mistral-small-latest; openai/gpt-oss-20b vs codestral-latest; openai/gpt-oss-20b vs mistral-code-latest; openai/gpt-oss-20b vs mistral-medium-latest; openai/gpt-oss-20b vs magistral-small-latest; openai/gpt-oss-20b vs ministral-14b-latest; openai/gpt-oss-20b vs ministral-8b-latest; openai/gpt-oss-20b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-20b vs nvidia/nemotron-3-super-120b-a12b:free; openai/gpt-oss-20b vs cohere/north-mini-code:free; openai/gpt-oss-20b vs minimax/minimax-m3:free; openai/gpt-oss-20b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; openai/gpt-oss-120b vs mistral-small-latest; openai/gpt-oss-120b vs codestral-latest; openai/gpt-oss-120b vs mistral-code-latest; openai/gpt-oss-120b vs mistral-medium-latest; openai/gpt-oss-120b vs magistral-small-latest; openai/gpt-oss-120b vs ministral-14b-latest; openai/gpt-oss-120b vs ministral-8b-latest; openai/gpt-oss-120b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-120b vs nvidia/nemotron-3-super-120b-a12b:free; openai/gpt-oss-120b vs cohere/north-mini-code:free; openai/gpt-oss-120b vs minimax/minimax-m3:free; openai/gpt-oss-120b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-small-latest vs codestral-latest; mistral-small-latest vs mistral-code-latest; mistral-small-latest vs mistral-medium-latest; mistral-small-latest vs magistral-small-latest; mistral-small-latest vs ministral-14b-latest; mistral-small-latest vs ministral-8b-latest; mistral-small-latest vs @cf/openai/gpt-oss-20b; mistral-small-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-small-latest vs cohere/north-mini-code:free; mistral-small-latest vs minimax/minimax-m3:free; mistral-small-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; codestral-latest vs mistral-code-latest; codestral-latest vs mistral-medium-latest; codestral-latest vs magistral-small-latest; codestral-latest vs ministral-14b-latest; codestral-latest vs ministral-8b-latest; codestral-latest vs @cf/openai/gpt-oss-20b; codestral-latest vs nvidia/nemotron-3-super-120b-a12b:free; codestral-latest vs cohere/north-mini-code:free; codestral-latest vs minimax/minimax-m3:free; codestral-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-code-latest vs mistral-medium-latest; mistral-code-latest vs magistral-small-latest; mistral-code-latest vs ministral-14b-latest; mistral-code-latest vs ministral-8b-latest; mistral-code-latest vs @cf/openai/gpt-oss-20b; mistral-code-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-code-latest vs cohere/north-mini-code:free; mistral-code-latest vs minimax/minimax-m3:free; mistral-code-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-medium-latest vs magistral-small-latest; mistral-medium-latest vs ministral-14b-latest; mistral-medium-latest vs ministral-8b-latest; mistral-medium-latest vs @cf/openai/gpt-oss-20b; mistral-medium-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-medium-latest vs cohere/north-mini-code:free; mistral-medium-latest vs minimax/minimax-m3:free; mistral-medium-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; magistral-small-latest vs ministral-14b-latest; magistral-small-latest vs ministral-8b-latest; magistral-small-latest vs @cf/openai/gpt-oss-20b; magistral-small-latest vs nvidia/nemotron-3-super-120b-a12b:free; magistral-small-latest vs cohere/north-mini-code:free; magistral-small-latest vs minimax/minimax-m3:free; magistral-small-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; ministral-14b-latest vs ministral-8b-latest; ministral-14b-latest vs @cf/openai/gpt-oss-20b; ministral-14b-latest vs nvidia/nemotron-3-super-120b-a12b:free; ministral-14b-latest vs cohere/north-mini-code:free; ministral-14b-latest vs minimax/minimax-m3:free; ministral-14b-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; ministral-8b-latest vs @cf/openai/gpt-oss-20b; ministral-8b-latest vs nvidia/nemotron-3-super-120b-a12b:free; ministral-8b-latest vs cohere/north-mini-code:free; ministral-8b-latest vs minimax/minimax-m3:free; ministral-8b-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; @cf/openai/gpt-oss-20b vs nvidia/nemotron-3-super-120b-a12b:free; @cf/openai/gpt-oss-20b vs cohere/north-mini-code:free; @cf/openai/gpt-oss-20b vs minimax/minimax-m3:free; @cf/openai/gpt-oss-20b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; nvidia/nemotron-3-super-120b-a12b:free vs cohere/north-mini-code:free; nvidia/nemotron-3-super-120b-a12b:free vs minimax/minimax-m3:free; nvidia/nemotron-3-super-120b-a12b:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; cohere/north-mini-code:free vs minimax/minimax-m3:free; cohere/north-mini-code:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; minimax/minimax-m3:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free. Below the 10-rollout floor for reading a rate as a rate: openai/gpt-oss-20b (1), openai/gpt-oss-120b (1), mistral-small-latest (0), codestral-latest (1), mistral-code-latest (1), mistral-medium-latest (0), magistral-small-latest (0), ministral-14b-latest (1), ministral-8b-latest (1), @cf/openai/gpt-oss-20b (1), nvidia/nemotron-3-super-120b-a12b:free (1), cohere/north-mini-code:free (1), minimax/minimax-m3:free (1), nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free (0) — 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	MISTRAL:MISTRAL-SMALL-LATEST	MISTRAL:CODESTRAL-LATEST	MISTRAL:MISTRAL-CODE-LATEST	MISTRAL:MISTRAL-MEDIUM-LATEST	MISTRAL:MAGISTRAL-SMALL-LATEST	
Inspect	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Annotate	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Validate	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Submit	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Graded	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Passed	1 / 1	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
groq:openai/gpt-oss-20b	Passed	Passed	1 / 1	0
groq:openai/gpt-oss-120b	Passed	Passed	1 / 1	0
mistral:mistral-small-latest	nothing	nothing	0 / 0	1
mistral:codestral-latest	Passed	Passed	1 / 1	0
mistral:mistral-code-latest	Passed	Passed	1 / 1	0
mistral:mistral-medium-latest	nothing	nothing	0 / 0	1
mistral:magistral-small-latest	nothing	nothing	0 / 0	1
mistral:ministral-14b-latest	Graded	Graded	0 / 1	0
mistral:ministral-8b-latest	Graded	Graded	0 / 1	0
cloudflare:@cf/openai/gpt-oss-20b	Passed	Passed	1 / 1	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	Graded	Graded	0 / 1	0
openrouter:cohere/north-mini-code:free	Graded	Graded	0 / 1	0
openrouter:minimax/minimax-m3:free	Passed	Passed	1 / 1	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	nothing	nothing	0 / 0	1

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

Annotation evidence

Modality: structured. **Rubric:** 1.0.

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

Submission ca520eff-c0a5-44e8-ae7-030142adc283

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 35e13207-d339-46f7-914e-c4d126d632a4

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 498af3be-8f81-4796-a913-b35bcf697d94

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 b9022701-959b-45bb-bf63-62f5ac576957

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 d91a7efb-333e-4d3e-a95e-539f29e34d52

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 33a5ee20-e744-43c0-8104-ffe80f003ae4

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 715ad3d1-a119-4a60-b124-4d5627f6e142

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 ca70ae26-cc53-4c91-ae73-ee4051aaa7f1

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 ddc4df0b-b41f-4576-97ea-94cc584c6b13

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

missing_annotation, missing_type

Disagreement

Agreement review Not measured

Submission 0640a190-81f0-4eb8-8ac9-dba23a2e9de8

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

Measured metrics

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

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

Validation error codes

None recorded

Disagreement

Agreement review Not measured

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

Quality axes

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

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
pairPrecision	recorded	N/A	1	–	–	
pairRecall	recorded	N/A	1	–	–	
pairF1	recorded	N/A	1	–	–	
pairAccuracy	recorded	N/A	1	–	–	
rowPrecision	recorded	N/A	0.9259	–	–	
rowRecall	recorded	N/A	0.9259	–	–	
rowF1	recorded	N/A	0.9259	–	–	
rowAccuracy	recorded	N/A	0.9259	–	–	

QUALITY SCORE (MEAN)

0.85

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	1	6 / 24	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	1	9 / 24	0	0 / 0	0	0	0	1
mistral:mistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:codestral-latest	1	6 / 24	0	0 / 0	0	0	0	0
mistral:mistral-code-latest	1	6 / 24	0	0 / 0	0	0	0	0
mistral:mistral-medium-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:magistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:ministral-14b-latest	1	24 / 24	0	0 / 0	0	0	0	4
mistral:ministral-8b-latest	1	8 / 24	0	0 / 0	0	0	0	0
cloudflare:@cf/openai/gpt-oss-20b	1	9 / 24	0	0 / 0	0	0	0	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	9 / 24	0	0 / 0	0	0	0	5
openrouter:cohere/north-mini-code:free	1	7 / 24	0	0 / 0	0	0	0	5
openrouter:minimax/minimax-m3:free	1	8 / 24	0	0 / 0	0	0	0	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	3 / 24	0	0 / 0	0	0	0	1

Each column is defined in Appendix D.

Hallucination & overclaim

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

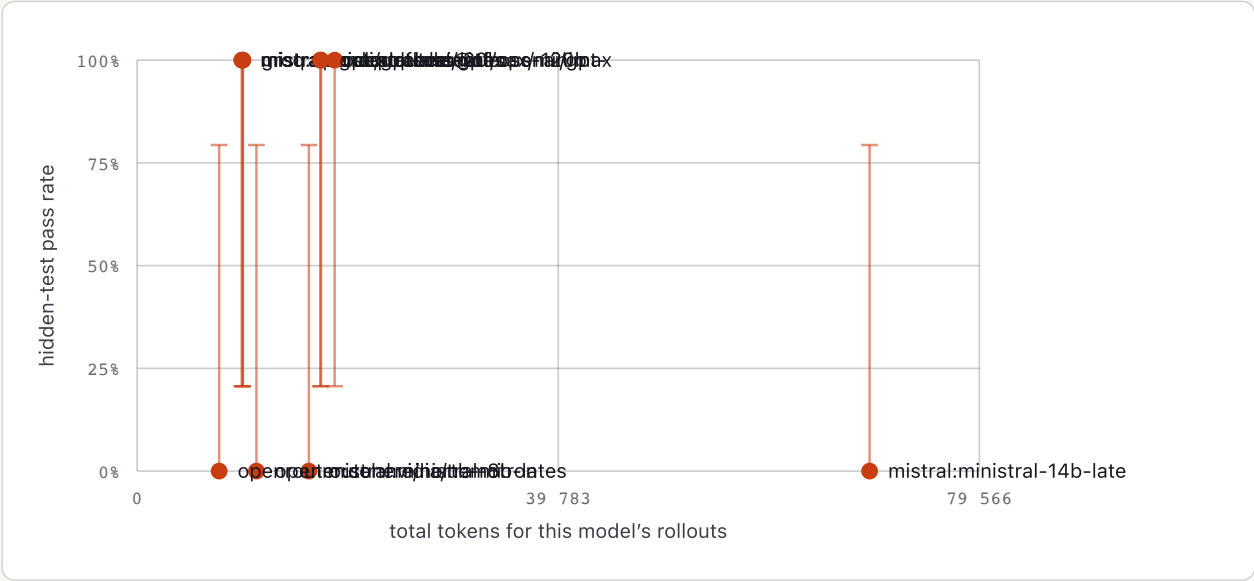
MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
groq:openai/gpt-oss-20b	1	0	none	0 / 1	0 / 1	0.0%	0
groq:openai/gpt-oss-120b	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-code-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-medium-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:magistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:ministral-14b-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:ministral-8b-latest	1	0	none	0 / 1	0 / 1	0.0%	0
cloudflare:@cf/openai/gpt-oss-20b	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:cohere/north-mini-code:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:minimax/minimax-m3:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 14 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	9 985	9 985	not billed	1 / 1	100.0%
groq:openai/gpt-oss-120b	1	17 309	17 309	not billed	1 / 1	100.0%
mistral:mistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:codestral-latest	1	9 990	9 990	not billed	1 / 1	100.0%
mistral:mistral-code-latest	1	9 879	9 879	not billed	1 / 1	100.0%
mistral:mistral-medium-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:magistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:ministral-14b-latest	1	69 188	69 188	not billed	0 / 1	0.0%
mistral:ministral-8b-latest	1	16 227	16 227	not billed	0 / 1	0.0%
cloudflare:@cf/openai/gpt-oss-20b	1	18 668	18 668	not billed	1 / 1	100.0%
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	11 277	11 277	not billed	0 / 1	0.0%
openrouter:cohere/north-mini-code:free	1	7 756	7 756	not billed	0 / 1	0.0%
openrouter:minimax/minimax-m3:free	1	17 386	17 386	not billed	1 / 1	100.0%
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	PASS
@cf/openai/gpt-oss-20b, codestral-latest, minimax/minimax-m3:free, mistral-code-latest, openai/gpt-oss-120b, openai/gpt-oss-20b passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
6 of 10 graded rollouts completed the applicable annotation workflow and were accepted by the independent annotation grader. This demonstrates the task is solvable under the published limits.	
F-01 · derived from the recorded run data	

F-02	LANE / INFRA
model_api_failure × 4 (magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free)	
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	ATTENTION
submitted_failed × 2 (ministral-14b-latest, ministral-8b-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Submitted annotations rejected by the independent grader.	
F-03 · derived from the recorded run data	

F-04	ATTENTION
incomplete_submission × 1 (nvidia/nemotron-3-super-120b-a12b:free)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
No description recorded for this class.	
F-04 · derived from the recorded run data	

F-05	ATTENTION
invalid_tool_loop × 1 (cohere/north-mini-code:free)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Repeatedly called tools that do not exist in this exam.	
F-05 · derived from the recorded run data	

Recommendations

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

R-01 PROBLEM The model endpoint failed to answer. Not a verdict about the model's ability. AFFECTED LANES magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free WHAT WE WOULD LOOK AT FIRST The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.
R-02 PROBLEM Submitted annotations rejected by the independent grader. AFFECTED LANES 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-03 PROBLEM Repeatedly called tools that do not exist in this exam. AFFECTED LANES cohere/north-mini-code:free 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-20260907T082138Z

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

4eb1c3548b8b094ded1a52d98f0179131be2555b821d9eea7b5fc519b0f684ce

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

SOURCE RECORD FILE

24b0c7c76a24db1e6373ba1dd43cbe5735b858af171f85a15e7b2c33fe59346f

sha256 of `fr-20260907-7c027ac8.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

91a0626344cddf180232e03f2e1354c04e21be8e7437214eb1609ae45e196de1

sha256 of `fr-20260907-7c027ac8.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

b56d36124274c44cfd31e8366e63c4eeb191ba35d2dd9a021cddc05f952819e9

sha256 of `fr-20260907-7c027ac8.pdf`.

RUN EVIDENCE BUNDLE

7c027ac8dcbe5e38bedd98cd43042409aa3849528f9efc7cde513adc2067134b

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
ca520eff	openai/gpt-oss-20b	0	submitted_passed	yes	Passed	pass	1	9 985	74.2s
35e13207	openai/gpt-oss-120b	0	submitted_passed	yes	Passed	pass	1	17 309	134.8s
4b1037c6	mistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.0s
498af3be	codestral-latest	0	submitted_passed	yes	Passed	pass	1	9 990	10.0s
b9022701	mistral-code-latest	0	submitted_passed	yes	Passed	pass	1	9 879	9.5s
3f6a408f	mistral-medium-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.9s
c556bdb6	magistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.2s
d91a7efb	ministral-14b-latest	0	submitted_failed	yes	Graded	fail	9	69 188	78.7s
33a5ee20	ministral-8b-latest	0	submitted_failed	yes	Graded	fail	1	16 227	20.9s
715ad3d1	@cf/openai/gpt-oss-20b	0	submitted_passed	yes	Passed	pass	1	18 668	27.7s
ca70ae26	nvidia/nemotron-3-super-120b-a12b:free	0	invalid_tool_loop	no	Graded	fail	1	11 277	111.8s
ddc4df0b	cohere/north-mini-code:free	0	invalid_tool_loop	no	Graded	fail	0	7 756	29.7s
0640a190	minimax/minimax-m3:free	0	submitted_passed	yes	Passed	pass	1	17 386	38.5s
97eecf9c	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	model_api_failure	no	Not graded — lane error	not graded	0	2 657	72.2s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	4	The model endpoint failed to answer. Not a verdict about the model's ability.
submitted_failed	2	Submitted annotations rejected by the independent grader.
incomplete_submission	1	No description recorded for this class.
invalid_tool_loop	1	Repeatedly called tools that do not exist in this exam.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST
7c027ac8dcbe5e38bedd98cd43
042409aa3849528f9efc7cde51
3adc2067134b

RUN ID
live-
20260907T082138Z

STARTED
2026-09-
07T08:21:38.363254+
00:00

FINISHED
2026-09-
07T08:38:05.395855
+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	bf8c69878d0999fc1bf2f5e9b1ef4280d6287acb28802ac64b1e5a3874f3691d
rollouts/0640a190-81f0-4eb8-8ac9-dba23a2e9de8.json	1b350335b8da944c770d75e4fc3c140275633d772b1b48510c4314b5af507c35
rollouts/33a5ee20-e744-43c0-8104-ffe80f003ae4.json	236056ac459b7db77f3a9a157f0b432d28c67881b924e8aa9845009cb98aa94d
rollouts/35e13207-d339-46f7-914e-c4d126d632a4.json	ac3ac1ec11b73edff1743d4d0f11fa9badd2c270f6f86cdc1ebefc02c202676e
rollouts/3f6a408f-4702-4663-afa1-6d55199953a6.json	c3368488acb2ff3321011131dd6edfb64ee52a03fe4541099194df566ec321a9
rollouts/498af3be-8f81-4796-a913-b35bcf697d94.json	34cd5d6aea2a5949d81ec9a6e1c85173e164f7fc209585210358ee4fdb5c454b
rollouts/4b1037c6-2090-4042-9eea-21363eab8aca.json	bb56b0b9f0207e8c67217444f5457ed0a6a6638892ba26962490162f2cb877f3
rollouts/715ad3d1-a119-4a60-b124-4d5627f6e142.json	2d8b688bc66a76142e658cd2e632b0cd18c614155134b8e1db15231d1135bbd8
rollouts/97eecf9c-dc83-4fa7-97c4-59b37fc653aa.json	b69b0f38d66ed5366a1d82ff978f44aff7dec6234e14c265eb60acaf09615c9f
rollouts/b9022701-959b-45bb-bf63-62f5ac576957.json	df791c2f2916c3217ae113dffac1de1a70e670fc40b2ec69e5a81b58d1ef35b
rollouts/c556bdb6-5897-44fa-8706-44b5e1c84622.json	44f880599111cb8a7c15d85ee3f9e06cfa5bbba908ab977a5877473966d2a94c
rollouts/ca520eff-c0a5-44e8-ae7-030142adc283.json	6d9cf8b0501c6a63ff4ce4acfd4a52772f1849889a56be674dfb8a191e4b2a93
rollouts/ca70ae26-cc53-4c91-ae73-ee4051aaa7f1.json	c90190931de6e8b351c0d6781a3ed8e9bbb9b6cbc0753b09296b24b66ecf10a6
rollouts/d91a7efb-333e-4d3e-a95e-539f29e34d52.json	7a45b53a6080f08baeb72f75cd5355d0ac3bbf392841f425a7906f5b98594547
rollouts/ddc4df0b-b41f-4576-97ea-94cc584c6b13.json	42df16fb0a617acac2c04d0e009bc480a9c0c52c767781aa3d6f5007ff05cba1
spec	7383db7f736bfa82c531cf0eb4190174be8c419b41d6ecf03edc17c86fbd2c38
summary	f69aa34275539b482516a478c7b486217c96a6f5947dc2cbd41b41d000476499

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.

ca520eff · 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.

35e13207 · groq:openai/gpt-oss-120b · 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	invalid_tool_call	(none)	failed	107
6	write_file	submission.json	ok	64
7	run_tests	—	ok	160
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (42 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.

4b1037c6 · mistral:mistral-small-latest · seed 0 · model_api_failure

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

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

498af3be · mistral:codestral-latest · seed 0 · submitted_passed

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

STEP	TOOL	PATH	RESULT	CHARS
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.

b9022701 · 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.

3f6a408f · mistral:mistral-medium-latest · seed 0 · model_api_failure

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

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

c556bdb6 · mistral:magistral-small-latest · seed 0 · model_api_failure

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

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

d91a7efb · 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	assets/records.json	ok	295
3	run_tests	—	ok	423
4	edit_file	submission.json	failed	29
5	write_file	submission.json	ok	64
6	run_tests	—	ok	268
7	run_tests	—	ok	268
8	edit_file	submission.json	ok	53

STEP	TOOL	PATH	RESULT	CHARS
9	write_file	submission.json	ok	64
10	write_file	submission.json	ok	64
11	write_file	submission.json	ok	64
12	run_tests	—	ok	268
13	write_file	submission.json	ok	64
14	run_tests	—	ok	257
15	write_file	submission.json	ok	92
16	run_tests	—	ok	354
17	read_file	annotation-task.json	ok	3 290
18	invalid_tool_call	(none)	failed	107
19	write_file	submission.json	ok	64
20	run_tests	—	ok	390
21	write_file	submission.json	ok	64
22	run_tests	—	ok	160
23	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.

33a5ee20 · mistral:ministral-8b-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	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.

715ad3d1 · 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	edit_file	submission.json	failed	29
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.

ca70ae26 · openrouter:nvidia/nemotron-3-super-120b-a12b:free · seed 0 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	invalid_tool_call	(none)	failed	107
2	read_file	TASK.md	ok	520
3	read_file	annotation-task.json	ok	3 290
4	read_file	submission.json	ok	74
5	invalid_tool_call	(none)	failed	107
6	write_file	submission.json	ok	64
7	invalid_tool_call	(none)	failed	107
8	invalid_tool_call	(none)	failed	107

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.

ddc4df0b · openrouter:cohere/north-mini-code:free · seed 0 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	invalid_tool_call	(none)	failed	107
2	read_file	—	failed	35
3	invalid_tool_call	(none)	failed	107
4	invalid_tool_call	(none)	failed	107
5	read_file	TASK.md	ok	520
6	invalid_tool_call	(none)	failed	107

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.

0640a190 · openrouter:minimax/minimax-m3:free · seed 0 · submitted_passed

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

Submission diff withheld by the public-disclosure policy (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.

97eecf9c · openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	81
1	invalid_tool_call	(none)	failed	107
2	read_file	TASK.md	ok	520

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.

12 Attempt record 1 of 3 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-0626d700 · record digest 1bef7c5673108536... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

14

CERTIFIED PASSES

3

of 9 graded rollouts

SUBMITTED

8 of 9

graded rollouts that called submit

NOT GRADED — LANE ERRORS

5

of 14 attempts

Purpose

This report presents a controlled evaluation of 14 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

3 of 9

graded rollouts passed the independent hidden tests and were certified — 5 of 14 attempts not graded (lane errors)

3 of 9 graded rollouts passed. @cf/openai/gpt-oss-20b, nvidia/nemotron-3-super-120b-a12b:free, openai/gpt-oss-120b 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, nvidia/nemotron-3-super-120b-a12b:free, openai/gpt-oss-120b passed the independent hidden tests
- F-02 model_api_failure × 5 (magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free, openai/gpt-oss-20b)
- F-03 submitted_failed × 5 (codestral-latest, minimax/minimax-m3:free, ministral-14b-latest, ministral-8b-latest, mistral-code-latest)
- F-04 invalid_tool_loop × 1 (cohere/north-mini-code:free)
- F-05 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: 1027.1s.

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

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.text-labeling-1	0.1.0	annotation	14 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

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

Evaluation configuration

REPORT	fr-20260907-0626d700
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.text-labeling-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:21:38	Evaluation run started
2026-09-07 08:38:45	Evaluation run finished
2026-09-07 08:38:45	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
mistral-small-latest	mistral	VVDex-configured	VVDex free-quota
codestral-latest	mistral	VVDex-configured	VVDex free-quota
mistral-code-latest	mistral	VVDex-configured	VVDex free-quota
mistral-medium-latest	mistral	VVDex-configured	VVDex free-quota
magistral-small-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
@cf/openai/gpt-oss-20b	cloudflare	VVDex-configured	VVDex free-quota
nvidia/nemotron-3-super-120b-a12b:free	openrouter	VVDex-configured	VVDex free-quota
cohere/north-mini-code:free	openrouter	VVDex-configured	VVDex free-quota
minimax/minimax-m3:free	openrouter	VVDex-configured	VVDex free-quota
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	openrouter	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: 14. Graded (evaluable) rollouts: 9. Not graded — lane errors: 5. 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	1	0	not graded	0	1	—	—	nothing
openai/gpt-oss-120b openai/gpt-oss-120b	1	1	1 / 1	0	0	1 / 1	64.7s	Passed
mistral-small-latest mistral-small-latest	1	0	not graded	0	1	—	—	nothing
codestral-latest codestral-latest	1	1	0 / 1	1	0	1 / 1	16.4s	Graded
mistral-code-latest mistral-code-latest	1	1	0 / 1	1	0	1 / 1	12.7s	Graded
mistral-medium-latest mistral-medium-latest	1	0	not graded	0	1	—	—	nothing
magistral-small-latest magistral-small-latest	1	0	not graded	0	1	—	—	nothing
ministral-14b-latest ministral-14b-latest	1	1	0 / 1	1	0	1 / 1	37.7s	Graded
ministral-8b-latest ministral-8b-latest	1	1	0 / 1	1	0	1 / 1	20.0s	Graded
@cf/openai/gpt-oss-20b @cf/openai/gpt-oss-20b	1	1	1 / 1	0	0	1 / 1	29.4s	Passed
nvidia/nemotron-3-super-120b-a12b:free nvidia/nemotron-3-super-120b-a12b:free	1	1	1 / 1	0	0	1 / 1	111.7s	Passed
cohere/north-mini-code:free cohere/north-mini-code:free	1	1	0 / 1	1	0	0 / 1	37.9s	Graded
minimax/minimax-m3:free minimax/minimax-m3:free	1	1	0 / 1	1	0	1 / 1	43.1s	Graded
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	not graded	0	1	—	—	nothing

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
openai/gpt-oss-20b	1	0	1	0	not recorded	[0%, 100%]	0	—
openai/gpt-oss-120b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
mistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
mistral-code-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
mistral-medium-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
magistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0	—
ministral-14b-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
ministral-8b-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
@cf/openai/gpt-oss-20b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
nvidia/nemotron-3-super-120b-a12b:free	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cohere/north-mini-code:free	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
minimax/minimax-m3:free	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	1	0	not recorded	[0%, 100%]	0	—

Interval and pass@k methods are stated in full in Appendix D. pass@k is shown for k = 1; 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	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
openai/gpt-oss-120b	100.0%	[21%, 100%]	12 178	not billed	64.7s	0.0%	none
mistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
codestral-latest	0.0%	[0%, 79%]	14 700	not billed	16.4s	0.0%	submitted_failed
mistral-code-latest	0.0%	[0%, 79%]	13 555	not billed	12.7s	0.0%	submitted_failed
mistral-medium-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
magistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
ministral-14b-latest	0.0%	[0%, 79%]	19 950	not billed	37.7s	0.0%	submitted_failed
ministral-8b-latest	0.0%	[0%, 79%]	14 356	not billed	20.0s	0.0%	submitted_failed
@cf/openai/gpt-oss-20b	100.0%	[21%, 100%]	10 789	not billed	29.4s	0.0%	none
nvidia/nemotron-3-super-120b-a12b:free	100.0%	[21%, 100%]	19 161	not billed	111.7s	0.0%	none
cohere/north-mini-code:free	0.0%	[0%, 79%]	6 563	not billed	37.9s	0.0%	invalid_tool_loop
minimax/minimax-m3:free	0.0%	[0%, 79%]	15 661	not billed	43.1s	0.0%	submitted_failed
nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure

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 mistral-small-latest; openai/gpt-oss-20b vs codestral-latest; openai/gpt-oss-20b vs mistral-code-latest; openai/gpt-oss-20b vs mistral-medium-latest; openai/gpt-oss-20b vs magistral-small-latest; openai/gpt-oss-20b vs ministral-14b-latest; openai/gpt-oss-20b vs ministral-8b-latest; openai/gpt-oss-20b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-20b vs nvidia/nemotron-3-super-120b-a12b:free; openai/gpt-oss-20b vs cohere/north-mini-code:free; openai/gpt-oss-20b vs minimax/minimax-m3:free; openai/gpt-oss-20b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; openai/gpt-oss-120b vs mistral-small-latest; openai/gpt-oss-120b vs codestral-latest; openai/gpt-oss-120b vs mistral-code-latest; openai/gpt-oss-120b vs mistral-medium-latest; openai/gpt-oss-120b vs magistral-small-latest; openai/gpt-oss-120b vs ministral-14b-latest; openai/gpt-oss-120b vs ministral-8b-latest; openai/gpt-oss-120b vs @cf/openai/gpt-oss-20b; openai/gpt-oss-120b vs nvidia/nemotron-3-super-120b-a12b:free; openai/gpt-oss-120b vs cohere/north-mini-code:free; openai/gpt-oss-120b vs minimax/minimax-m3:free; openai/gpt-oss-120b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-small-latest vs codestral-latest; mistral-small-latest vs mistral-code-latest; mistral-small-latest vs mistral-medium-latest; mistral-small-latest vs magistral-small-latest; mistral-small-latest vs ministral-14b-latest; mistral-small-latest vs ministral-8b-latest; mistral-small-latest vs @cf/openai/gpt-oss-20b; mistral-small-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-small-latest vs cohere/north-mini-code:free; mistral-small-latest vs minimax/minimax-m3:free; mistral-small-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; codestral-latest vs mistral-code-latest; codestral-latest vs mistral-medium-latest; codestral-latest vs magistral-small-latest; codestral-latest vs ministral-14b-latest; codestral-latest vs ministral-8b-latest; codestral-latest vs @cf/openai/gpt-oss-20b; codestral-latest vs nvidia/nemotron-3-super-120b-a12b:free; codestral-latest vs cohere/north-mini-code:free; codestral-latest vs minimax/minimax-m3:free; codestral-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-code-latest vs mistral-medium-latest; mistral-code-latest vs magistral-small-latest; mistral-code-latest vs ministral-14b-latest; mistral-code-latest vs ministral-8b-latest; mistral-code-latest vs @cf/openai/gpt-oss-20b; mistral-code-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-code-latest vs cohere/north-mini-code:free; mistral-code-latest vs minimax/minimax-m3:free; mistral-code-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; mistral-medium-latest vs magistral-small-latest; mistral-medium-latest vs ministral-14b-latest; mistral-medium-latest vs ministral-8b-latest; mistral-medium-latest vs @cf/openai/gpt-oss-20b; mistral-medium-latest vs nvidia/nemotron-3-super-120b-a12b:free; mistral-medium-latest vs cohere/north-mini-code:free; mistral-medium-latest vs minimax/minimax-m3:free; mistral-medium-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; magistral-small-latest vs ministral-14b-latest; magistral-small-latest vs ministral-8b-latest; magistral-small-latest vs @cf/openai/gpt-oss-20b; magistral-small-latest vs nvidia/nemotron-3-super-120b-a12b:free; magistral-small-latest vs cohere/north-mini-code:free; magistral-small-latest vs minimax/minimax-m3:free; magistral-small-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; ministral-14b-latest vs ministral-8b-latest; ministral-14b-latest vs @cf/openai/gpt-oss-20b; ministral-14b-latest vs nvidia/nemotron-3-super-120b-a12b:free; ministral-14b-latest vs cohere/north-mini-code:free; ministral-14b-latest vs minimax/minimax-m3:free; ministral-14b-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; ministral-8b-latest vs @cf/openai/gpt-oss-20b; ministral-8b-latest vs nvidia/nemotron-3-super-120b-a12b:free; ministral-8b-latest vs cohere/north-mini-code:free; ministral-8b-latest vs minimax/minimax-m3:free; ministral-8b-latest vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; @cf/openai/gpt-oss-20b vs nvidia/nemotron-3-super-120b-a12b:free; @cf/openai/gpt-oss-20b vs cohere/north-mini-code:free; @cf/openai/gpt-oss-20b vs minimax/minimax-m3:free; @cf/openai/gpt-oss-20b vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; nvidia/nemotron-3-super-120b-a12b:free vs cohere/north-mini-code:free; nvidia/nemotron-3-super-120b-a12b:free vs minimax/minimax-m3:free; nvidia/nemotron-3-super-120b-a12b:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; cohere/north-mini-code:free vs minimax/minimax-m3:free; cohere/north-mini-code:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free; minimax/minimax-m3:free vs nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free. Below the 10-rollout floor for reading a rate as a rate: openai/gpt-oss-20b (0), openai/gpt-oss-120b (1), mistral-small-latest (0), codestral-latest (1), mistral-code-latest (1), mistral-medium-latest (0), magistral-small-latest (0), ministral-14b-latest (1), ministral-8b-latest (1), @cf/openai/gpt-oss-20b (1), nvidia/nemotron-3-super-120b-a12b:free (1), cohere/north-mini-code:free (1), minimax/minimax-m3:free (1), nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free (0) — 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	MISTRAL:MISTRAL-SMALL-LATEST	MISTRAL:CODESTRAL-LATEST	MISTRAL:MISTRAL-CODE-LATEST	MISTRAL:MISTRAL-MEDIUM-LATEST	MISTRAL:MAGISTRAL-SMALL-LATEST	
Inspect	0 / 0	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Annotate	0 / 0	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Validate	0 / 0	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Submit	0 / 0	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Graded	0 / 0	1 / 1	0 / 0	1 / 1	1 / 1	0 / 0	0 / 0	
Passed	0 / 0	1 / 1	0 / 0	0 / 1	0 / 1	0 / 0	0 / 0	

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
groq:openai/gpt-oss-20b	nothing	nothing	0 / 0	1
groq:openai/gpt-oss-120b	Passed	Passed	1 / 1	0
mistral:mistral-small-latest	nothing	nothing	0 / 0	1
mistral:codestral-latest	Graded	Graded	0 / 1	0
mistral:mistral-code-latest	Graded	Graded	0 / 1	0
mistral:mistral-medium-latest	nothing	nothing	0 / 0	1
mistral:magistral-small-latest	nothing	nothing	0 / 0	1
mistral:ministral-14b-latest	Graded	Graded	0 / 1	0
mistral:ministral-8b-latest	Graded	Graded	0 / 1	0
cloudflare:@cf/openai/gpt-oss-20b	Passed	Passed	1 / 1	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	Passed	Passed	1 / 1	0
openrouter:cohere/north-mini-code:free	Graded	Graded	0 / 1	0
openrouter:minimax/minimax-m3:free	Graded	Graded	0 / 1	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	nothing	nothing	0 / 0	1

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

Annotation evidence

Modality: text. **Rubric:** 1.0.

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

Submission 42b7ee3c-a8e2-4183-be69-1c20c11c1004

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 56132f8d-a6a2-45e4-9a31-102b15935f56

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 a656c3b2-ef8e-4873-941a-744290c67e0e

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 7b94e943-816a-49c3-b221-354933910d9f

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 edeb008a-8c57-4e34-8418-1ee62bf0741e

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

Submission 8d1a8d13-fb5b-443e-abca-5728ecd2715a

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 036d30d1-e5e2-4db7-b695-8adb91c41d27

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 8ddb2019-cc57-42af-b0b9-42a6eb6a6c9b

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

missing_annotation, missing_type

Disagreement

Agreement review

Not measured

Submission 3b701009-2b32-4b16-9ed3-ecbcbe7dbdf6

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

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	0.6875	–	–	
spanRecall	recorded	N/A	0.6875	–	–	
spanF1	recorded	N/A	0.6875	–	–	
spanOverlapIoU	recorded	N/A	0.7132	–	–	
spanExactF1	recorded	N/A	0.625	–	–	

QUALITY SCORE (MEAN)

0.7189

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	1	4 / 24	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	1	7 / 24	0	0 / 0	0	0	0	0
mistral:mistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:codestral-latest	1	9 / 24	0	0 / 0	0	0	0	1
mistral:mistral-code-latest	1	8 / 24	0	0 / 0	0	0	0	0
mistral:mistral-medium-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:magistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:ministral-14b-latest	1	10 / 24	0	0 / 0	0	0	0	2
mistral:ministral-8b-latest	1	8 / 24	0	0 / 0	0	0	0	0
cloudflare:@cf/openai/gpt-oss-20b	1	6 / 24	0	0 / 0	0	0	0	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	11 / 24	0	0 / 0	0	0	0	3
openrouter:cohere/north-mini-code:free	1	6 / 24	0	0 / 0	0	0	0	7
openrouter:minimax/minimax-m3:free	1	8 / 24	0	0 / 0	0	0	0	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	7 / 24	0	0 / 0	0	0	0	3

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.

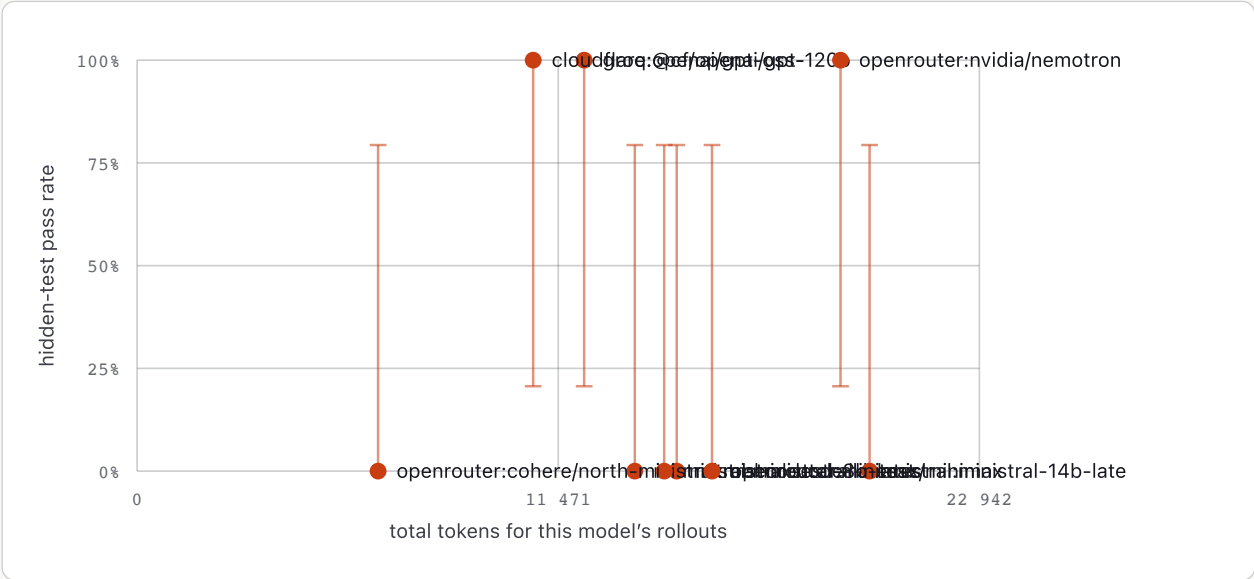
MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
groq:openai/gpt-oss-20b	1	0	none	0 / 1	0 / 1	0.0%	0
groq:openai/gpt-oss-120b	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-code-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-medium-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:magistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:ministral-14b-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:ministral-8b-latest	1	0	none	0 / 1	0 / 1	0.0%	0
cloudflare:@cf/openai/gpt-oss-20b	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:cohere/north-mini-code:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:minimax/minimax-m3:free	1	0	none	0 / 1	0 / 1	0.0%	0
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 14 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	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
groq:openai/gpt-oss-120b	1	12 178	12 178	not billed	1 / 1	100.0%
mistral:mistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:codestral-latest	1	14 700	14 700	not billed	0 / 1	0.0%
mistral:mistral-code-latest	1	13 555	13 555	not billed	0 / 1	0.0%
mistral:mistral-medium-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:magistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:ministral-14b-latest	1	19 950	19 950	not billed	0 / 1	0.0%
mistral:ministral-8b-latest	1	14 356	14 356	not billed	0 / 1	0.0%
cloudflare:@cf/openai/gpt-oss-20b	1	10 789	10 789	not billed	1 / 1	100.0%
openrouter:nvidia/nemotron-3-super-120b-a12b:free	1	19 161	19 161	not billed	1 / 1	100.0%
openrouter:cohere/north-mini-code:free	1	6 563	6 563	not billed	0 / 1	0.0%
openrouter:minimax/minimax-m3:free	1	15 661	15 661	not billed	0 / 1	0.0%
openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	PASS
@cf/openai/gpt-oss-20b, nvidia/nemotron-3-super-120b-a12b:free, openai/gpt-oss-120b passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
3 of 9 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 × 5 (magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free, 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	ATTENTION
submitted_failed × 5 (codestral-latest, minimax/minimax-m3:free, ministral-14b-latest, ministral-8b-latest, mistral-code-latest)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Submitted annotations rejected by the independent grader.	
F-03 · derived from the recorded run data	

F-04	ATTENTION
invalid_tool_loop × 1 (cohere/north-mini-code:free)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
Repeatedly called tools that do not exist in this exam.	
F-04 · derived from the recorded run data	

F-05	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R1 api-failure share 5/14 — 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-05 · derived from the recorded run data	

Recommendations

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

R-01 PROBLEM The model endpoint failed to answer. Not a verdict about the model's ability. AFFECTED LANES magistral-small-latest, mistral-medium-latest, mistral-small-latest, nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free, 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.
R-02 PROBLEM Submitted annotations rejected by the independent grader. AFFECTED LANES codestral-latest, minimax/minimax-m3:free, ministral-14b-latest, ministral-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-03 PROBLEM Repeatedly called tools that do not exist in this exam. AFFECTED LANES cohere/north-mini-code:free 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-20260907T082138Z

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

1bef7c56731085366e1774b7eaef75d81f12d3b2afeb9276f028c23ae441b421

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

SOURCE RECORD FILE

1142bfb3b75c59bc4fc7ae4078f19b2846ef3c065353ea689a6aa817f837b2ed

sha256 of `fr-20260907-0626d700.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

4d7cc9a195376c4d191f80eff126be5705aed4e56c3e1b444b700275424e735f

sha256 of `fr-20260907-0626d700.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

f3d2c60ac36e748897255ba7add774e3dd0049ccc7a62a541bffa4422d3c340

sha256 of `fr-20260907-0626d700.pdf`.

RUN EVIDENCE BUNDLE

0626d7007bdbc125459aa7403e52f28e9ea31c0ca7acf353d9b2641087497801

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
00d2f76f	openai/gpt-oss-20b	0	model_api_failure	no	Not graded — lane error	not graded	1	5 764	50.8s
42b7ee3c	openai/gpt-oss-120b	0	submitted_passed	yes	Passed	pass	1	12 178	64.7s
2711c9ab	mistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	57.7s
56132f8d	codestral-latest	0	submitted_failed	yes	Graded	fail	1	14 700	16.4s
a656c3b2	mistral-code-latest	0	submitted_failed	yes	Graded	fail	1	13 555	12.7s
751ee6e6	mistral-medium-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	57.8s
a46488ea	magistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	60.1s
7b94e943	ministral-14b-latest	0	submitted_failed	yes	Graded	fail	1	19 950	37.7s
edeb008a	ministral-8b-latest	0	submitted_failed	yes	Graded	fail	1	14 356	20.0s
8d1a8d13	@cf/openai/gpt-oss-20b	0	submitted_passed	yes	Passed	pass	1	10 789	29.4s
036d30d1	nvidia/nemotron-3-super-120b-a12b:free	0	submitted_passed	yes	Passed	pass	1	19 161	111.7s
8ddb2019	cohere/north-mini-code:free	0	invalid_tool_loop	no	Graded	fail	0	6 563	37.9s
3b701009	minimax/minimax-m3:free	0	submitted_failed	yes	Graded	fail	1	15 661	43.1s
9cbd2bdc	nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free	0	model_api_failure	no	Not graded — lane error	not graded	1	15 964	223.4s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	5	The model endpoint failed to answer. Not a verdict about the model's ability.
submitted_failed	5	Submitted annotations rejected by the independent grader.
invalid_tool_loop	1	Repeatedly called tools that do not exist in this exam.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

0626d7007bdbc125459aa7403e52f28e9ea31c0ca7acf353d9b2641087497801

RUN ID

live-20260907T082138Z

STARTED

2026-09-07T08:21:38.363252+00:00

FINISHED

2026-09-07T08:38:45.468309+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	0c00718c0612db9fdbcb3d42301e4616ab37abfb4e47be21ef9bb7d95171a6f71
rollouts/00d2f76f-977f-44d4-811a-b2b5ce417744.json	370bc274329ea2f616774a0cbb7c7c2d1ceca9a66b44430e5ef3d65ca3617dc9
rollouts/036d30d1-e5e2-4db7-b695-8adb91c41d27.json	8868a40bd037f3093acc5a526e8df344d774a161e1e9fbc4eb39919764faec3c
rollouts/2711c9ab-0622-4847-912a-97fb3b2383f8.json	65e151dcfee517570aebf177c3af904b1dfe5cf08717b94640fc09351e1e386a
rollouts/3b701009-2b32-4b16-9ed3-ecbcbe7dbdf6.json	c911887c932be47ca93105a4e94aa6c34d5d27d2d9b654154f2f226b0d06add
rollouts/42b7ee3c-a8e2-4183-be69-1c20c11c1004.json	0d1c1c260f221e2a9979b06b93d70834b562262d39858860e904f0b0f5fa21ff
rollouts/56132f8d-a6a2-45e4-9a31-102b15935f56.json	0b282db14e6ea24ace92c2aad60c1cddb357a7f8418c9572f2e38939fdbed935
rollouts/751ee6e6-0aae-492c-81be-3f1e1bd44b10.json	a8d8d0bf828cb54b93656df1fac668b8107535c1142efdc18f3cead8f535e077
rollouts/7b94e943-816a-49c3-b221-354933910d9f.json	2b72e3c02f015623c22d80f0074792e83e5710d9294decdec5cb077955c76aa0b
rollouts/8d1a8d13-fb5b-443e-abca-5728ecd2715a.json	6dcda11c467f3a5c72ed40b22bea94abc98741f3cb82b35e92b0915b608e36ce
rollouts/8ddb2019-cc57-42af-b0b9-42a6eb6a6c9b.json	d936846104372fa5dbadaca3142818312d0813c4f8407bd7109ae155020c16af
rollouts/9cbd2bdc-98bd-445f-823b-2f940eb7002d.json	e6da0ec3a6e3ae82b18f9b12c7649b6e98b7b4ab3ca0d4d0e953975541c1458c
rollouts/a46488ea-6823-49b7-9a77-afddccf1d98e.json	be2b16086038165ff704719ca4d12db38ee2618165acb38e65318bc6d9d9843c
rollouts/a656c3b2-ef8e-4873-941a-744290c67e0e.json	6ecc7306b5554229be6b8e731e071f96a2d887d433abb21223a18df054c7665c
rollouts/edeb008a-8c57-4e34-8418-1ee62bf0741e.json	12000e9e9630c7c92589d444a724826af9472a71868336c7991e32f0c04d3540
spec	203a1718aba753379bdafdc5ae71b661dd5fba952cfb41b4f36ea27c6b163242
summary	67a910d4c49e6e70a8964af0da3c5fe112c8cf60774f59e6831ea6997fde91d8

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.

00d2f76f · 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.

42b7ee3c · 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	2 826
3	read_file	submission.json	ok	74
4	edit_file	submission.json	ok	53
5	run_tests	—	ok	160
6	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.

2711c9ab · mistral:mistral-small-latest · seed 0 · model_api_failure

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

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

56132f8d · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	TASK.md	ok	409
1	list_files	.	ok	80
2	invalid_tool_call	(none)	failed	107
3	read_file	annotation-task.json	ok	2 826
4	read_file	assets/message.txt	ok	69

STEP	TOOL	PATH	RESULT	CHARS
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 (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.

a656c3b2 · mistral:mistral-code-latest · seed 0 · 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.

751ee6e6 · mistral:mistral-medium-latest · seed 0 · model_api_failure

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

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

a46488ea · mistral:magistral-small-latest · seed 0 · model_api_failure

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

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

7b94e943 · 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	assets/message.txt	ok	69
3	read_file	annotation-task.json	ok	2 826
4	run_tests	—	ok	442
5	run_tests	—	ok	442
6	run_tests	—	ok	442
7	write_file	submission.json	ok	64

STEP	TOOL	PATH	RESULT	CHARS
8	run_tests	—	ok	160
9	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.

edeb008a · 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	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.

8d1a8d13 · 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	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.

036d30d1 · openrouter:nvidia/nemotron-3-super-120b-a12b:free · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	invalid_tool_call	(none)	failed	107
1	list_files	.	ok	80
2	read_file	TASK.md	ok	409
3	read_file	annotation-task.json	ok	2 826
4	invalid_tool_call	(none)	failed	107
5	read_file	assets/message.txt	ok	69
6	invalid_tool_call	(none)	failed	107
7	read_file	submission.json	ok	74
8	write_file	submission.json	ok	64
9	run_tests	—	ok	160
10	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.

8ddb2019 · openrouter:cohere/north-mini-code:free · seed 0 · invalid_tool_loop

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	80
1	read_file	—	failed	35
2	invalid_tool_call	(none)	failed	107
3	invalid_tool_call	(none)	failed	107
4	invalid_tool_call	(none)	failed	107
5	invalid_tool_call	(none)	failed	107

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.

3b701009 · openrouter:minimax/minimax-m3:free · 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	read_file	submission.json	ok	74
5	write_file	submission.json	ok	64
6	run_tests	—	ok	160
7	submit	—	ok	85

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

9cbd2bdc · openrouter:nvidia/nemotron-3-nano-omni-30b-a3b-reasoning:free · seed 0 · model_api_failure

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

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

vvdex.annotation.text-labeling-1 · record fr-20260907-5b08f93e · record digest b8a5d79aaefa53e7... · harness SUSPECT · {"breached": 1, "unverified": 0}

Executive summary

MODELS TESTED

3

CERTIFIED PASSES

2

of 3 graded rollouts

SUBMITTED

3 of 3

graded rollouts that called submit

NOT GRADED — LANE ERRORS

0

of 3 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

2 of 3

graded rollouts passed the independent hidden tests and were certified

2 of 3 graded rollouts passed. cli/claude-sonnet, cli/codex completed the applicable annotation workflow and were accepted by the independent annotation grader. **Low-N:** 3 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was isolation_breach — .

Key findings

- F-01

cli/claude-sonnet, cli/codex passed the independent hidden tests
- F-02

isolation_breach × 1 (cli/cursor-grok)
- F-03

1 rollout(s) acted outside the exam channel — isolation breach
- F-04

The harness suspected itself during this run
- F-05

Statistical floor — 3 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 162.2s.

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 14520s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.text-labeling-1	0.1.0	annotation	3 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

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

Evaluation configuration

REPORT	fr-20260907-5b08f93e
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.text-labeling-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:45:02	Evaluation run started
2026-09-07 08:47:44	Evaluation run finished
2026-09-07 08:47:44	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Cursor Agent CLI (Grok 4.6 Fast, subscription)	cli	local runtime	operator subscription, flat — not billed per token

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

Limits

RUNTIME docker	STEP LIMIT 24	WALL-CLOCK LIMIT 14520s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 3. Graded (evaluable) rollouts: 3. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Codex CLI (GPT, subscription) cli/codex	1	1	1 / 1	0	0	1 / 1	53.7s	Passed
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	1	1	1 / 1	0	0	1 / 1	25.0s	Passed
Cursor Agent CLI (Grok 4.6 Fast, subscription) cli/cursor-grok	1	1	0 / 1	1	0	1 / 1	81.4s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1
cli/codex	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/cursor-grok	1	1	0	0	0.0%	[0%, 79%]	0	0.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/codex	100.0%	[21%, 100%]	n/a	n/a	53.7s	0.0%	none
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	25.0s	0.0%	none
cli/cursor-grok	0.0%	[0%, 79%]	n/a	n/a	81.4s	0.0%	isolation_breach

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

This is not a leaderboard. Not separated (intervals overlap): cli/codex vs cli/claude-sonnet; cli/codex vs cli/cursor-grok; cli/claude-sonnet vs cli/cursor-grok. Below the 10-rollout floor for reading a rate as a rate: cli/codex (1), cli/claude-sonnet (1), cli/cursor-grok (1) — 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	CLI:CLI/CODEX	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CURSOR-GROK
Inspect	1 / 1	1 / 1	1 / 1
Annotate	1 / 1	1 / 1	1 / 1
Validate	1 / 1	1 / 1	1 / 1
Submit	1 / 1	1 / 1	1 / 1
Graded	1 / 1	1 / 1	1 / 1
Passed	1 / 1	1 / 1	0 / 1

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
cli:cli/codex	Passed	Passed	1 / 1	0
cli:cli/claude-sonnet	Passed	Passed	1 / 1	0
cli:cli/cursor-grok	Graded	Graded	0 / 1	0

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

Annotation evidence

Modality: text. **Rubric:** 1.0.

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

Submission 2757d4d7-6e2d-4143-9398-99d53fc87f75

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 f545d679-7f7e-45d3-8b62-5fd258d4b75a

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 6e28ca24-2c21-4e3d-8c5b-84a576a5b1b8

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

Measured metrics

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

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

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

Validation error codes

None recorded

Disagreement

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

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

Quality axes

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

AXIS	GATE	ROLLOUTS PASSING	MEASURED	BUDGET	WEIGHT	WHAT IT MEASURES
classificationPrecision	recorded	N/A	1	–	–	
classificationRecall	recorded	N/A	1	–	–	
classificationF1	recorded	N/A	1	–	–	
classificationAccuracy	recorded	N/A	1	–	–	
spanPrecision	recorded	N/A	1	–	–	
spanRecall	recorded	N/A	1	–	–	
spanF1	recorded	N/A	1	–	–	
spanOverlapIoU	recorded	N/A	1	–	–	
spanExactF1	recorded	N/A	1	–	–	

QUALITY SCORE (MEAN)

1

Weighted reading over the axes. It is not the reward.

REWARD COMPOSITION

Withheld from the public report.

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
cli:cli/codex	1	6 / 24	0	0 / 0	0	0	0	0
cli:cli/claude-sonnet	1	5 / 24	0	0 / 0	0	0	0	0
cli:cli/cursor-grok	1	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
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/cursor-grok	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 3 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

F-01	PASS
cli/claude-sonnet, cli/codex passed the independent hidden tests	
Severity: PASS	Evidence: Model outcomes · Stage funnel · Appendix A
2 of 3 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
isolation_breach × 1 (cli/cursor-grok)	
Severity: ATTENTION	Evidence: Appendix A · Appendix C
No description recorded for this class.	
F-02 · derived from the recorded run data	

F-03	ATTENTION
1 rollout(s) acted outside the exam channel — isolation breach	
Severity: ATTENTION	Evidence: Appendix A
The taker used its own tools, memory or network rather than the exam's actions. Affected lane(s): cli/cursor-grok. 1 pass(es) were withdrawn: a grader exit code earned outside the channel is not a model result, and this report does not present it as one.	
F-03 · derived from the recorded run data	

F-04	NOTE
The harness suspected itself during this run	
Severity: NOTE	Evidence: Appendix A · Appendix C
Self-suspicion rules fired: R4 containment breach in 1/3 rollouts (lanes ['cli/cursor-grok']) — takers acted outside the exam channel; those rows are not model results. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-04 · derived from the recorded run data	

F-05	NOTE
Statistical floor — 3 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-05 · derived from the recorded run data	

Recommendations

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

R-01 For anyone reading the percentages

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

Verification

RUN ID

live-20260907T084502Z

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

b8a5d79aaefa53e7c5be413210df59f8d994e411b9f71ccaaea2b78a135503e0

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

SOURCE RECORD FILE

52aed65d01a8415920803abedb956aefecc3c30b1169e56a1e3845a15beea862

sha256 of `fr-20260907-5b08f93e.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

21bbcc25d7e52b3517825ea99469a89a8b9bac49e67c42c387224c3116c6f849

sha256 of `fr-20260907-5b08f93e.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

c00d265e3f3c048067947904d2d58dfdeeb87813c514d144f85e58d2023aae8a

sha256 of `fr-20260907-5b08f93e.pdf`.

RUN EVIDENCE BUNDLE

5b08f93eac7cb058076d685f1d1e18730a1071fab31cc040658e7b61a33197d9

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

CERTIFIED EVIDENCE SEAL

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

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

Measurement history

Previous run on this exam: `fr-20260907-f7b0a094`. 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
2757d4d7	cli/codex	0	submitted_passed	yes	Passed	pass	1	n/a	53.7s
f545d679	cli/claude-sonnet	0	submitted_passed	yes	Passed	pass	1	n/a	25.0s
6e28ca24	cli/cursor-grok	0	submitted_passed	yes	Graded	fail	1	n/a	81.4s

Failure taxonomy

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

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST
5b08f93eac7cb058076d685f1d
1e18730a1071fab31cc040658e
7b61a33197d9

RUN ID
live-
20260907T084502Z

STARTED
2026-09-
07T08:45:02.533066
+00:00

FINISHED
2026-09-
07T08:47:44.709376+
00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	1b4a8f7b60b6342a624714b49414dc34386e823a617e9e61fedb22ea04fbed31
rollouts/2757d4d7-6e2d-4143-9398-99d53fc87f75.json	ff3f67be4ce672900f8918942492af2607d90bf1483ed8c12c799c58537cb605
rollouts/6e28ca24-2c21-4e3d-8c5b-84a576a5b1b8.json	178b1dfe9a88026a24cef545ce8238008d43535780d065033b701b39373459b7
rollouts/f545d679-7f7e-45d3-8b62-5fd258d4b75a.json	c98e0c10d45f2033b13a176186ae2335e63d5129b64d01a36d5cf31854046e95
spec	95a97d1da59a1700b2e7f310fc496445c84effb6bb46b2b127d26eaffd5d40bf
summary	58f70e4b543e1ba5768a339a229636013797dd2af5994bafbed8577c43d6b946

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.

2757d4d7 · cli:cli/codex · seed 0 · submitted_passed

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

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

f545d679 · cli:cli/claude-sonnet · seed 0 · submitted_passed

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

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

6e28ca24 · cli:cli/cursor-grok · seed 0 · submitted_passed

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

14 Attempt record 3 of 3 · text-labeling-1

vvdex.annotation.text-labeling-1 · record fr-20260907-f7b0a094 · record digest cb855a9a633a7a61... · harness SUSPECT · containment clean

Executive summary

MODELS TESTED

3

CERTIFIED PASSES

0

of 0 graded rollouts

SUBMITTED

0 of 0

graded rollouts that called submit

NOT GRADED — LANE ERRORS

3

of 3 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

0 of 0

graded rollouts passed the independent hidden tests and were certified — 3 of 3 attempts not graded (lane errors)

0 of 0 graded rollouts passed the hidden tests. The exam's certification establishes the task is winnable, so every failure below is a finding about a lane, not about the instrument. **Low-N:** 3 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was model_api_failure — the model endpoint failed to answer. not a verdict about the model's ability.

Key findings

- F-01 No evaluated lane passed the hidden tests in this run
- F-02 model_api_failure × 3 (magistral-small-latest, mistral-medium-latest, mistral-small-latest)
- F-03 The harness suspected itself during this run
- F-04 Statistical floor — 3 rollout(s) is below the 10-rollout rate floor

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 216.5s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the asset, annotate against the declared rubric, validate, submit, and be accepted by the independent annotation grader — within 24 tool steps and 2280s of wall clock, at 1 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

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

What was measured

EXAM	VERSION	FAMILY	ROLLOUTS
vvdex.annotation.text-labeling-1	0.1.0	annotation	3 (1 per model)

Where the defect came from

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

Certification evidence

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

FROZEN BASELINE

FAIL

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

→

GOLD / REFERENCE

PASS

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

GRADER CONTROLS

19 / 19

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

ATTACK PROBES

9 / 9 blocked

0 trivial exploit(s) found.

SEALED EVIDENCE

verified

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

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

CERTIFICATION BUNDLE DIGEST

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadc7b4ed5ace282

EXAM FINGERPRINT

e1abe51398b2d950a3415678...

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

Evaluation configuration

REPORT	fr-20260907-f7b0a094
ISSUED	2026-09-07
SUBJECT EXAM	vvdex.annotation.text-labeling-1 · v0.1.0
FAMILY	annotation
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: recorded and passing (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-06 00:01:41	Exam evidence sealed (before any model saw the exam)
2026-09-07 08:43:12	Evaluation run started
2026-09-07 08:46:48	Evaluation run finished
2026-09-07 08:46:48	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
mistral-small-latest	mistral	VVDex-configured	VVDex free-quota
mistral-medium-latest	mistral	VVDex-configured	VVDex free-quota
magistral-small-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: 3. Graded (evaluable) rollouts: 0. Not graded — lane errors: 3. 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
mistral-small-latest mistral-small-latest	1	0	not graded	0	1	—	—	nothing
mistral-medium-latest mistral-medium-latest	1	0	not graded	0	1	—	—	nothing
magistral-small-latest magistral-small-latest	1	0	not graded	0	1	—	—	nothing

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS
mistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0
mistral-medium-latest	1	0	1	0	not recorded	[0%, 100%]	0
magistral-small-latest	1	0	1	0	not recorded	[0%, 100%]	0

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
mistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
mistral-medium-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure
magistral-small-latest	not recorded	[0%, 100%]	n/a	n/a	0ms	0.0%	model_api_failure

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): mistral-small-latest vs mistral-medium-latest; mistral-small-latest vs magistral-small-latest; mistral-medium-latest vs magistral-small-latest. Below the 10-rollout floor for reading a rate as a rate: mistral-small-latest (0), mistral-medium-latest (0), magistral-small-latest (0) — 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	MISTRAL:MISTRAL-SMALL-LATEST	MISTRAL:MISTRAL-MEDIUM-LATEST	MISTRAL:MAGISTRAL-SMALL-LATEST
Inspect	0 / 0	0 / 0	0 / 0
Annotate	0 / 0	0 / 0	0 / 0
Validate	0 / 0	0 / 0	0 / 0
Submit	0 / 0	0 / 0	0 / 0
Graded	0 / 0	0 / 0	0 / 0
Passed	0 / 0	0 / 0	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	ANNOTATION QA PASSES	NOT GRADED
mistral:mistral-small-latest	nothing	nothing	0 / 0	1
mistral:mistral-medium-latest	nothing	nothing	0 / 0	1
mistral:magistral-small-latest	nothing	nothing	0 / 0	1

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

Annotation evidence

Modality: text. **Rubric:** 1.0.

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

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

Behavioural analysis

Long-horizon behaviour

Counted off the recorded trace of each rollout. A dead end is a step that moved nothing: a re-read of a slice already returned, a repeat of the previous action, or a call to a tool this exam does not have.

MODEL	N	STEPS USED	RAN OUT OF STEPS	COMMANDS	CHANGED THE TREE	REFUSED (OVER BUDGET)	REVISITS	DEAD ENDS / ROLLOUT
mistral:mistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:mistral-medium-latest	1	0 / 24	0	0 / 0	0	0	0	0
mistral:magistral-small-latest	1	0 / 24	0	0 / 0	0	0	0	0

Each column is defined in Appendix D.

Hallucination & overclaim

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

MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
mistral:mistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:mistral-medium-latest	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:magistral-small-latest	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 3 rollout(s). No lane re-read material it had already been given.

Every term above is defined in Appendix D.

Efficiency & telemetry

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

MODEL	ROLLOUTS	TOKENS	TOKENS / ROLLOUT	COST	PASSES	RATE
mistral:mistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:mistral-medium-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%
mistral:magistral-small-latest	0	n/a	n/a	n/a – telemetry unavailable	0 / 0	0.0%

Findings

Every finding below is derived from the recorded data of this run — none is an opinion.

Severity: PASS a lane succeeded · ATTENTION a capability gap · LANE / INFRA infrastructure, not capability · NOTE a property of the measurement.

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

F-02	LANE / INFRA
model_api_failure × 3 (magistral-small-latest, mistral-medium-latest, mistral-small-latest)	
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 3/3 — most rollouts never got a working completion. Treat the affected lanes' numbers as measurements of the harness, not the models, until the incident is resolved.	
F-03 · derived from the recorded run data	

F-04	NOTE
Statistical floor — 3 rollout(s) is below the 10-rollout rate floor	
Severity: NOTE	Evidence: Appendix D
The counts are exact; the percentages are anecdotes with decimal points. Per-rollout outcomes, not rates, are the reliable reading of this run.	
F-04 · derived from the recorded run data	

Recommendations

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

R-01 PROBLEM

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

AFFECTED LANES

magistral-small-latest, mistral-medium-latest, mistral-small-latest

WHAT WE WOULD LOOK AT FIRST

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

R-02 For anyone reading the percentages

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

Verification

RUN ID

live-20260907T084312Z

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

cb855a9a633a7a6172ac2dbab698cf69f3de94376eca921aad3451254036599e

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

SOURCE RECORD FILE

703e6708db3c937afb0a22e5fb207efe478b79d3fa1ea096e2705b1c39aa314a

sha256 of `fr-20260907-f7b0a094.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

28aee31617332f84506423edfa3f96f1c98f11b087806d0ee29e9b9d4e148a7f

sha256 of `fr-20260907-f7b0a094.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

963b69d80cd78db6134ab95e59087e34caa5eb8e68eb28780064ad05826aa9bf

sha256 of `fr-20260907-f7b0a094.pdf`.

RUN EVIDENCE BUNDLE

f7b0a094d79e8b6ba9e6bc66d0c722cc49f83b6909865b163b8ce3720d8b812d

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

CERTIFIED EVIDENCE SEAL

0f97b9d0560aeb01635959125ac2e126ab1103e0087113deadcb7b4ed5ace282

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

Measurement history

Previous run on this exam: `fr-20260907-0626d700`. 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
70c5477b	mistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.6s
a08c5660	mistral-medium-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.8s
7d42e188	magistral-small-latest	0	model_api_failure	no	Not graded — lane error	not graded	0	0	58.3s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	3	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

f7b0a094d79e8b6ba9e6bc66d0c722cc49f83b6909865b163b8ce3720d8b812d

RUN ID

live-20260907T084312Z

STARTED

2026-09-07T08:43:12.056792+00:00

FINISHED

2026-09-07T08:46:48.530409+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	c6cea04066b733b3666917da6521fa4532faf40579b4fc94512ed4dbe17320ac
rollouts/70c5477b-7001-4cf7-9c69-b87331003531.json	f67bf75c373fd1946faa5c1f29f5e9450b9a47774bc4acc716c12fbcc9cb4a15
rollouts/7d42e188-ea6e-41e9-b703-bb16b2690aa5.json	4123a90d145742acf0b034b5df18e684a36d6768c6608c50a630a68b37a3f302
rollouts/a08c5660-d922-44e4-8963-7d56229f73e3.json	d1d1d4e56e50a01be730dc2ac7e4c071788714e1bf75cdb47f1e66538a4927ce
spec	da420bc256223c33bff204afa1ec617b95860fd231a0edc3caac836698c7e717
summary	22ab2a7cc71e92a2e44f9ef357a34151c5f84df18bf297d200e35f57bde86dd1

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.

70c5477b · mistral:mistral-small-latest · seed 0 · model_api_failure

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

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

a08c5660 · mistral:mistral-medium-latest · seed 0 · model_api_failure

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

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

7d42e188 · mistral:magistral-small-latest · seed 0 · model_api_failure

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

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

15 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-20260907-15d51112` · `recordDigest` `d074d0dd0c2e7a855207c95b302bfc08bc7a092265803601f8a19c28ea6a2bb`
- `vvdex.annotation.structured-data-1` → `fr-20260907-29701401` · `recordDigest` `4d674e21c58427397f8b9b5918e587f9111427e72eaa2281bb7d0fcfa7bd7848`
- `vvdex.annotation.structured-data-1` → `fr-20260907-7c027ac8` · `recordDigest` `4eb1c3548b8b094ded1a52d98f0179131be2555b821d9eea7b5fc519b0f684ce`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-0626d700` · `recordDigest` `1bef7c56731085366e1774b7eae75d81f12d3b2afeb9276f028c23ae441b421`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-5b08f93e` · `recordDigest` `b8a5d79aaefa53e7c5be413210df59f8d994e411b9f71ccaea2b78a135503e0`
- `vvdex.annotation.text-labeling-1` → `fr-20260907-f7b0a094` · `recordDigest` `cb855a9a633a7a6172ac2dbab698cf69f3de94376eca921aad3451254036599e`

16 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.