Certified campaign — one rollout per lane

5 sealed evaluations · 16 graded rollouts · 8 certified passes

HISTORICAL — SUPERSEDED

CAMPAIGN REPORT

2026-09-01

Evaluation scope:

cli/claude-sonnet
cli/codex
codestral-latest
openai/gpt-oss-120b

Prepared by:

VVDex Forge — regenerated from verified sealed records only

`fc-8626f712e26f` · 5 record(s) · every value recomputable

Contents

01 Executive summary

Release status: HISTORICAL — superseded. A newer campaign carries these exams. This report is published so the earlier measurement stays checkable; its exams are NOT on the public featured index, and the current numbers are in the campaign the index names.

EXAMS	SEALED RECORDS	MODEL LANES	ATTEMPTS
5	5	4	20

GRADED ROLLOUTS	CERTIFIED PASSES	LANE ERRORS (NOT GRADED)	WITHHELD / INVALID
16	8	4	0

Purpose

This report presents a controlled evaluation of 4 model lane(s) against 5 fixed certified exams, one rollout per lane. For each exam, every lane received that exam'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 evaluation 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.

Outcome

8

certified pass(es) across 16 graded rollouts (20 attempts, 4 lane error(s) not graded) in 20 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-8626f712e26f · generated from 5 verified sealed record(s)

Key findings across the campaign

- `martinblech-xmltodict-issue-257` · `cli/claude-sonnet` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `martinblech-xmltodict-issue-257` · `cli/codex` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `martinblech-xmltodict-issue-257` · `codestral-latest` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `martinblech-xmltodict-issue-257` · `openai/gpt-oss-120b` · **NO RESULT** — 0 certified passes / 0 graded rollouts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `pytoolz-toolz-issue-602` · `cli/claude-sonnet` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `pytoolz-toolz-issue-602` · `cli/codex` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `pytoolz-toolz-issue-602` · `codestral-latest` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `pytoolz-toolz-issue-602` · `openai/gpt-oss-120b` · **NO RESULT** — 0 certified passes / 0 graded rollouts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `r1chardj0n3s-parse-issue-102` · `cli/claude-sonnet` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

- `r1chardj0n3s-parse-issue-102 · cli/codex` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `r1chardj0n3s-parse-issue-102 · codestral-latest` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `r1chardj0n3s-parse-issue-102 · openai/gpt-oss-120b` · **NO RESULT** — 0 certified passes / 0 graded rollouts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `grounded-rag-1 · cli/claude-sonnet` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `grounded-rag-1 · cli/codex` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `grounded-rag-1 · codestral-latest` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `grounded-rag-1 · openai/gpt-oss-120b` · **NO RESULT** — 0 certified passes / 0 graded rollouts; 0 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- `memory-fact-update-1 · cli/claude-sonnet` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `memory-fact-update-1 · cli/codex` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `memory-fact-update-1 · codestral-latest` · **ATTENTION** — 0 certified passes / 1 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- `memory-fact-update-1 · openai/gpt-oss-120b` · **PASS** — 1 certified passes / 1 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	SWE · XMLTODICT-257 PUBLIC.SWE.MARTINBLECH-XMLTODICT-ISSUE-257	SWE · TOOLZ-602 PUBLIC.SWE.PYTOOLZ-TOOLZ-ISSUE-602	SWE · PARSE-102 PUBLIC.SWE.R1CHARDJ0N3S-PARSE-ISSUE-102	RAG VVDEX.KNOWLEDGE.GROUNDED-RAG-1	MEMORY VVDEX.KNOWLEDGE.MEMORY-FACT-UPDATE-1
cli/claude-sonnet	1/1	1/1	1/1	0/1	1/1
cli/codex	1/1	0/1	0/1	1/1	1/1
openai/gpt-oss-120b	0/0 + 1 lane error	0/0 + 1 lane error	0/0 + 1 lane error	0/0 + 1 lane error	1/1
codestral-latest	0/1	0/1	0/1	0/1	0/1

Each cell is *certified passes / graded rollouts* 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 five exam(s) drawn from three 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
Software repair (public OSS SWE)	`public.swe.martinblech-xmltodict-issue-257`	A real software-engineering task taken from a public open-source repository: the model is given the codebase and a written issue, and must change the code so that the behaviour described in the issue is fixed.
Software repair (public OSS SWE)	`public.swe.pytoolz-toolz-issue-602`	A real software-engineering task taken from a public open-source repository: the model is given the codebase and a written issue, and must change the code so that the behaviour described in the issue is fixed.
Software repair (public OSS SWE)	`public.swe.r1chardj0n3s-parse-issue-102`	A real software-engineering task taken from a public open-source repository: the model is given the codebase and a written issue, and must change the code so that the behaviour described in the issue is fixed.
Grounded retrieval (RAG)	`vvdex.knowledge.grounded-rag-1`	Grounded question answering over a small document collection.
Cross-session memory update	`vvdex.knowledge.memory-fact-update-1`	A state-and-memory task across sessions: a fact was recorded earlier, an authoritative source has since changed it, and the model must answer from the newer authoritative roster rather than the stale carried memory, and acknowledge the change.

04 Campaign aggregate across these five exams

MODEL LANE	CERTIFIED PASSES	GRADED ROLLOUTS	LANE ERRORS	WITHHELD	INVALID	CERTIFIED RATE (95% CI)
cli/claude-sonnet	**4**	5	0	0	0	—
cli/codex	**3**	5	0	0	0	—
openai/gpt-oss-120b	**1**	1	4	0	0	—
codestral-latest	**0**	5	0	0	0	—

Raw counts first. *Graded rollouts* = certified passes + model fails; lane errors, withheld and invalid cells are outside capability arithmetic in both directions. Denominators differ between lanes, so no rate is shown and none should be inferred — a lane with one countable attempt has one data point, not a percentage.

05 Harness status

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

06 Cells excluded from capability

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

EXAM	LANE	SEED	EXCLUSION	RECORDED REASON
martinblech-xmltodict-issue-257	`openai/gpt-oss-120b`	0	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): …
pytoolz-toolz-issue-602	`openai/gpt-oss-120b`	0	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): …
r1chardj0n3s-parse-issue-102	`openai/gpt-oss-120b`	0	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): …
grounded-rag-1	`openai/gpt-oss-120b`	0	lane error	HTTP 429: {"error":{"message":"Rate limit reached for model `openai/gpt-oss-120b` in organization [redacted] service tier `on_demand` on tokens per day (TPD): …

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
martinblech-xmltodict-issue-257	`public_source`	`Applied consistently to 1 sealed records.` public-source disclosure (ownership public_source declared by engine registry, upstream https://github.com/martinblech/xmltodict under MIT; family public_oss_swe): the submitted change is a patch against a public repository, so a redacted excerpt of it is published.
pytoolz-toolz-issue-602	`answer_key`	`Applied consistently to 1 sealed records.` answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family public_oss_swe): 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.
r1chardj0n3s-parse-issue-102	`answer_key`	`Applied consistently to 1 sealed records.` answer-key disclosure (ownership vvdex_proprietary — no declaration — proprietary by default; family public_oss_swe): 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.
grounded-rag-1	`answer_key`	`Applied consistently to 1 sealed records.` answer-key disclosure (ownership vvdex_proprietary — engine registry; family rag_grounding): 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.
memory-fact-update-1	`answer_key`	`Applied consistently to 1 sealed records.` answer-key disclosure (ownership vvdex_proprietary — engine registry; family long_term_memory): 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.

Exam revisions retired for future evaluation

- `vvdex.knowledge.grounded-rag-1 0.1.0` — Retired for future evaluation on 2026-09-02: post-run public disclosure of answer-bearing submission content. Historical sealed results on this revision are unaffected: the disclosure happened after those runs, so it could not have influenced them.
- `vvdex.knowledge.memory-fact-update-1 0.1.0` — Retired for future evaluation on 2026-09-02: post-run public disclosure of answer-bearing submission content. Historical sealed results on this revision are unaffected: the disclosure happened after those runs, so it could not have influenced them.

09 Exam · martinblech-xmltodict-issue-257

`public.swe.martinblech-xmltodict-issue-257 · record fr-20260901-a27316a3 · record digest 96f8b631c6d65f03… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
4	2 of 3 graded rollouts	3 of 3 graded rollouts that called submit	1 of 4 attempts

Purpose

This report presents a controlled evaluation of 4 model lane(s) against the certified exam `public.swe.martinblech-xmltodict-issue-257` (v0.1.0, family public_oss_swe). 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 — 1 of 4 attempts not graded (lane errors)

2 of 3 graded rollouts passed. cli/claude-sonnet, cli/codex completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. **Low-N:** 4 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	cli/claude-sonnet, cli/codex passed the independent hidden tests
F-02	model_api_failure × 1 (openai/gpt-oss-120b)
F-03	submitted_failed × 1 (codestral-latest)
F-04	Statistical floor — 4 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: 1398.7s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the workspace, produce a correct edit, verify it with the visible tests, and submit — within 40 tool steps and 9720s 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 `public.swe.martinblech-xmltodict-issue-257` 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
public.swe.martinblech-xmltodict-issue-257	0.1.0	public_oss_swe	4 (1 per model)

Where the defect came from

Source: martinblech/xmltodict (MIT) · issue https://github.com/martinblech/xmltodict/issues/257 · upstream fix https://github.com/martinblech/xmltodict/pull/421

The defect and its upstream fix are public and unmodified. The exam around them — the frozen parent tree, the independently written hidden suite, the control probes and the isolation — is VVDex's, and it is what this report measures.

Certification evidence

4 of 5 certification pillars are recorded in this package, loaded from the receipts sealed into the exam and matching this run's exam fingerprint. What is recorded is shown with its real value; what is not is named below and is not claimed.

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

7 / 7

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

ATTACK PROBES

10 / 10 blocked

0 trivial exploit(s) found.

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

Certification metadata is not included in this report package for:

- sealed evidence bundle

Benchmark results remain valid as run evidence; the pillar(s) above are not substantiated by this document and this report does not claim them.

What the package does carry: the exam fingerprint below is the contract digest this run was executed against — a holder of the certification bundle for this fingerprint can join the two independently.

EXAM FINGERPRINT

aa41977120e6ecdddb5c4fc603e8ab23cb272726a86750119969ff4dfc90e297

The contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260901-a27316a3
ISSUED	2026-09-01
SUBJECT EXAM	public.swe.martinblech-xmltodict-issue-257 · v0.1.0
FAMILY	public_oss_swe
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: partially recorded (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-01 22:35:32	Evaluation run started
2026-09-01 22:58:51	Evaluation run finished
2026-09-01 22:58:51	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
codestral-latest	mistral	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	40	9720s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 4. Graded (evaluable) rollouts: 3. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) `cli/claude-sonnet`	1	1	1 / 1	0	0	1 / 1	110.2s	Passed hidden tests
Codex CLI (GPT, subscription) `cli/codex`	1	1	1 / 1	0	0	1 / 1	40.1s	Passed hidden tests
codestral-latest `codestral-latest`	1	1	0 / 1	1	0	1 / 1	12.5s	Graded
openai/gpt-oss-120b `openai/gpt-oss-120b`	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
cli/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/codex	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
openai/gpt-oss-120b	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
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	110.2s	0.0%	none
cli/codex	100.0%	[21%, 100%]	n/a	n/a	40.1s	0.0%	none
codestral-latest	0.0%	[0%, 79%]	19 214	not billed	12.5s	0.0%	submitted_failed
openai/gpt-oss-120b	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): cli/claude-sonnet vs cli/codex; cli/claude-sonnet vs codestral-latest; cli/claude-sonnet vs openai/gpt-oss-120b; cli/codex vs codestral-latest; cli/codex vs openai/gpt-oss-120b; codestral-latest vs openai/gpt-oss-120b. Below the 10-rollout floor for reading a rate as a rate: cli/claude-sonnet (1), cli/codex (1), codestral-latest (1), openai/gpt-oss-120b (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

The funnel is the primary reading of a run. Where a model stops says more than a single pass percentage does, especially at low N.

STAGE	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CODEX	MISTRAL:CODESTRAL-LATEST	GROQ:OPENAI/GPT-OSS-120B
Inspected	1 / 1	1 / 1	1 / 1	0 / 0
Edited	1 / 1	1 / 1	1 / 1	0 / 0
Ran tests	1 / 1	1 / 1	1 / 1	0 / 0
Submitted	1 / 1	1 / 1	1 / 1	0 / 0
Graded	1 / 1	1 / 1	1 / 1	0 / 0
Passed hidden tests	1 / 1	1 / 1	0 / 1	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Passed hidden tests	Passed hidden tests	1 / 1	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	1 / 1	0
mistral:codestral-latest	Graded	Graded	0 / 1	0
groq:openai/gpt-oss-120b	nothing	nothing	0 / 0	1

Stage definitions are in Appendix D. The matrix counts evaluable rollouts; lane errors are shown beside it, not inside it. The stages are not strictly nested: a model can submit without editing, and the matrix shows that rather than hiding it. Each rollout's own deepest stage is in Appendix A.

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/claude-sonnet	1	5 / 40	0	0 / 0	0	0	0	0
cli:cli/codex	1	6 / 40	0	0 / 0	0	0	0	0
mistral:codestral-latest	1	5 / 40	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	1	1 / 40	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/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	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

No grounding failures and no false pass claims were detected across these 4 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
cli:cli/claude-sonnet	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
cli:cli/codex	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
mistral:codestral-latest	1	19 214	19 214	not billed	0 / 1	0.0%
groq:openai/gpt-oss-120b	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

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 full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. 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 × 1 (openai/gpt-oss-120b)

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 × 1 (codestral-latest)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted a change; the independent hidden tests still failed.

F-03 · derived from the recorded run data

F-04 NOTE

Statistical floor — 4 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

openai/gpt-oss-120b

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 a change; the independent hidden tests still failed.

AFFECTED LANES

codestral-latest

WHAT WE WOULD LOOK AT FIRST

The model completes the loop and produces a plausible near-miss — the named failing tests are the exact behavioral cases distinguishing it from the reference fix.

R–03

For anyone reading the percentages

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

Verification

RUN ID

live-20260901T223532Z

EXAM FINGERPRINT

aa41977120e6ecdddb5c4fc603e8ab23cb272726a86750119969ff4dfc90e297

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

96f8b631c6d65f03b21e4a9a94e86c82874e1506fac38c9a093d2f8c618dd2dc

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

SOURCE RECORD FILE

cac23cf090b72e918199a0968244efe9871320c3a2a5a2b9863e1a0963a4c5b0

sha256 of fr-20260901-a27316a3.json as written beside the standalone report.

SOURCE STANDALONE REPORT

b72719ef41d6a6d15f78c00ffcbfb2c6fa0d3646faf6593d72c87a23e4574880

sha256 of fr-20260901-a27316a3.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

fcc2bb4c0484075ca2aa80d21f3b7661b07703df0876c45915704e5b5ba172e4

sha256 of fr-20260901-a27316a3.pdf.

RUN EVIDENCE BUNDLE

a27316a347e66d960235e2f5d5880c5254a556000866fcb91eb0bb85ed4af5be

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

CERTIFIED EVIDENCE SEAL

not available in this package

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
65ba9fe6	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	110.2s
3b5cf81e	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	40.1s
1f03a051	codestral-latest	0	submitted_failed	yes	Graded	fail	1	19 214	12.5s
43ea93d0	openai/gpt-oss-120b	0	model_api_failure	no	Not graded — lane error	not graded	0	1 217	1214.5s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.
submitted_failed	1	Submitted a change; the independent hidden tests still failed.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
a27316a347e66d960235e2f5d5880c5254a556000866fcb91eb0bb85ed4af5be	live-20260901T223532Z	2026-09-01T22:35:32.913870+00:00	2026-09-01T22:58:51.597668+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	d8d74d56cb9e61d158de398f5b616841217b544c8385eade746f8a2783cb5904
rollouts/1f03a051-d5ed-46f8-b421-3b08eff2ec50.json	feec4f7afbf85a7a77a5105fd5afd24c305e1b771f56076460b4fcd55fa61eeb
rollouts/3b5cf81e-0754-435f-9d06-74d408b02016.json	b735926f1987113807c6bd342fb658cabf8830f4772fda0fac7e2112b07576ed
rollouts/43ea93d0-f7b9-401b-913d-5094ad34ade8.json	4468f2fa8ec26459d52c476cfd8f9acb4069d67601c0233cda2e808807866f45
rollouts/65ba9fe6-0de3-4733-9f6a-e444ce5e84e4.json	a02851de9dec2c935ad19eb603f209279f0f63bd748076438ab84a57a85d09f5
spec	b32fc5d6978a10421f3e522abfe530c59b462251a2e6c58e69b2eb6ea2a84a30
summary	584de4b70cfc06667ca9462857141a67d4928902d13aa3c3a60544442b20e681

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, and an excerpt of the diff it produced. What a tool returned is not here — only which tool, on which of the model's own paths, whether it succeeded, and how many characters came back. 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, because printing it would put the hidden tree's layout in a file that leaves this machine. The diff excerpt is shown because this exam's disclosure class is public-source: the change is a patch against a public repository under the licence named in the exam definition.

Disclosure class applied to this exam: public-source disclosure (ownership public_source declared by engine registry, upstream https://github.com/martinblech/xmltodict under MIT; family public_oss_swe): the submitted change is a patch against a public repository, so a redacted excerpt of it is published.

65ba9fe6 · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	xmltodict.py	ok	12 166
1	read_file	xmltodict.py	ok	11 706
2	edit_file	xmltodict.py	ok	50
3	run_tests	–	ok	183
4	submit	–	ok	85

```
--- a/xmltodict.py
+++ b/xmltodict.py
@@ -107,21 +107,41 @@
         # without attaching it back to its parent. This avoids accumulating all
         # streamed items in memory when using item_depth > 0.
         if len(self.path) == self.item_depth:
+            data = (None if not self.data
+                    else self.cdata_separator.join(self.data))
             item = self.item
-            if item is None:
-                item = (None if not self.data
-                        else self.cdata_separator.join(self.data))
-
-            should_continue = self.item_callback(self.path, item)
-            if not should_continue:
-                raise ParsingInterrupted
-            # Reset state for the parent context without keeping a reference to
-            # the emitted item.
             if self.stack:
                 self.item, self.data = self.stack.pop()
             else:
                 self.item = None
                 self.data = []
+            # Apply the same processing as the non-streaming branch so that
+            # the value handed to the callback equals what the element would
+            # have in a full (non-streaming) parse.
+            if self.strip_whitespace and data:
+                data = data.strip() or None
+            if data and self._should_force_cdata(name, data) and item is None:
+                item = self.dict_constructor()
+            if item is not None:
+                if data:
+                    self.push_data(item, self.cdata_key, data)
+                # Determine the value that would be stored in the parent
+                # (applying postprocessor the same way push_data would).
+                if self.postprocessor is not None:
+                    result = self.postprocessor(self.path, name, item)
+                    callback_value = item if result is None else result[1]
```

```
+                else:
+                    callback_value = item
```

Showing 40 of 53 diff line(s); truncated.

3b5cf81e · cli:cli/codex · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	ISSUE.md	ok	1 282
2	read_file	xmltodict.py	ok	12 166
3	edit_file	xmltodict.py	ok	50
4	run_tests	–	ok	183
5	submit	–	ok	85

```
--- a/xmltodict.py
+++ b/xmltodict.py
@@ -107,10 +107,27 @@
         # without attaching it back to its parent. This avoids accumulating all
         # streamed items in memory when using item_depth > 0.
         if len(self.path) == self.item_depth:
+            data = (None if not self.data
+                    else self.cdata_separator.join(self.data))
             item = self.item
-            if item is None:
-                item = (None if not self.data
-                        else self.cdata_separator.join(self.data))
+            if self.strip_whitespace and data:
+                data = data.strip() or None
+            if data and self._should_force_cdata(name, data) and item is None:
+                item = self.dict_constructor()
+            if item is not None:
+                if data:
+                    self.push_data(item, self.cdata_key, data)
+            else:
+                item = data
+
+            # Apply the same element-level postprocessing and force-list logic
+            # as push_data() would when attaching this value to its parent, but
+            # keep it in a temporary container so streaming does not retain it.
+            emitted = self.push_data(None, name, item)
+            if emitted is None:
+                item = None
+            else:
+                item = next(iter(emitted.values()))

             should_continue = self.item_callback(self.path, item)
             if not should_continue:
```

Showing 33 of 33 diff line(s).

1f03a051 · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 282
1	read_file	xmltodict.py	ok	12 166
2	edit_file	xmltodict.py	ok	50

STEP	TOOL	PATH	RESULT	CHARS
3	run_tests	–	ok	183
4	submit	–	ok	85

```
--- a/xmltodict.py
+++ b/xmltodict.py
@@ -111,6 +111,16 @@
             if item is None:
                 item = (None if not self.data
                         else self.cdata_separator.join(self.data))
+            else:
+                if self.data:
+                    data = self.cdata_separator.join(self.data)
+                    if self.strip_whitespace:
+                        data = data.strip() or None
+                    if data:
+                        if self._should_force_cdata(self.path[-1][0], data):
+                            item[self.cdata_key] = data
+                        else:
+                            item[self.cdata_key] = data

             should_continue = self.item_callback(self.path, item)
             if not should_continue:
```

Showing 19 of 19 diff line(s).

43ea93d0 · groq:openai/gpt-oss-120b · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39

This rollout changed no file, so there is no diff to show.

10 Exam · pytoolz-toolz-issue-602

`public.swe.pytoolz-toolz-issue-602 · record fr-20260901-88c3cdf1 · record digest f49a31336e4865c1… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
4	1 of 3 graded rollouts	3 of 3 graded rollouts that called submit	1 of 4 attempts

Purpose

This report presents a controlled evaluation of 4 model lane(s) against the certified exam `public.swe.pytoolz-toolz-issue-602` (v0.1.0, family public_oss_swe). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 3

graded rollouts passed the independent hidden tests and were certified — 1 of 4 attempts not graded (lane errors)

1 of 3 graded rollouts passed. cli/claude-sonnet completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. **Low-N:** 4 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was `submitted_failed` — submitted a change; the independent hidden tests still failed.

Key findings

F-01	cli/claude-sonnet passed the independent hidden tests
F-02	submitted_failed × 2 (cli/codex, codestral-latest)
F-03	model_api_failure × 1 (openai/gpt-oss-120b)
F-04	Statistical floor — 4 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: 1315.4s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the workspace, produce a correct edit, verify it with the visible tests, and submit — within 40 tool steps and 9720s 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 `public.swe.pytoolz-toolz-issue-602` 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
public.swe.pytoolz-toolz-issue-602	0.1.0	public_oss_swe	4 (1 per model)

Where the defect came from

Source: pytoolz/toolz (BSD-3-Clause) · issue https://github.com/pytoolz/toolz/issues/602 · upstream fix https://github.com/pytoolz/toolz/pull/603

The defect and its upstream fix are public and unmodified. The exam around them — the frozen parent tree, the independently written hidden suite, the control probes and the isolation — is VVDex's, and it is what this report measures.

Certification evidence

4 of 5 certification pillars are recorded in this package, loaded from the receipts sealed into the exam and matching this run's exam fingerprint. What is recorded is shown with its real value; what is not is named below and is not claimed.

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

7 / 7

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

ATTACK PROBES

10 / 10 blocked

0 trivial exploit(s) found.

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

Certification metadata is not included in this report package for:

- sealed evidence bundle

Benchmark results remain valid as run evidence; the pillar(s) above are not substantiated by this document and this report does not claim them.

What the package does carry: the exam fingerprint below is the contract digest this run was executed against — a holder of the certification bundle for this fingerprint can join the two independently.

EXAM FINGERPRINT

e8312ce98b82558f9a1c5593e7d431f409fb1024b3013b7dbd934f1398a8937f

The contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260901-88c3cdf1
ISSUED	2026-09-01
SUBJECT EXAM	public.swe.pytoolz-toolz-issue-602 · v0.1.0
FAMILY	public_oss_swe
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: partially recorded (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-01 21:49:07	Evaluation run started
2026-09-01 22:11:03	Evaluation run finished
2026-09-01 22:11:03	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
codestral-latest	mistral	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	40	9720s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 4. Graded (evaluable) rollouts: 3. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) `cli/claude-sonnet`	1	1	1 / 1	0	0	1 / 1	93.2s	Passed hidden tests
Codex CLI (GPT, subscription) `cli/codex`	1	1	0 / 1	1	0	1 / 1	54.8s	Graded
codestral-latest `codestral-latest`	1	1	0 / 1	1	0	1 / 1	16.7s	Graded
openai/gpt-oss-120b `openai/gpt-oss-120b`	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
cli/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/codex	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
openai/gpt-oss-120b	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
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	93.2s	0.0%	none
cli/codex	0.0%	[0%, 79%]	n/a	n/a	54.8s	0.0%	submitted_failed
codestral-latest	0.0%	[0%, 79%]	40 054	not billed	16.7s	0.0%	submitted_failed
openai/gpt-oss-120b	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): cli/claude-sonnet vs cli/codex; cli/claude-sonnet vs codestral-latest; cli/claude-sonnet vs openai/gpt-oss-120b; cli/codex vs codestral-latest; cli/codex vs openai/gpt-oss-120b; codestral-latest vs openai/gpt-oss-120b. Below the 10-rollout floor for reading a rate as a rate: cli/claude-sonnet (1), cli/codex (1), codestral-latest (1), openai/gpt-oss-120b (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

The funnel is the primary reading of a run. Where a model stops says more than a single pass percentage does, especially at low N.

STAGE	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CODEX	MISTRAL:CODESTRAL-LATEST	GROQ:OPENAI/GPT-OSS-120B
Inspected	1 / 1	1 / 1	1 / 1	0 / 0
Edited	1 / 1	1 / 1	0 / 1	0 / 0
Ran tests	1 / 1	1 / 1	1 / 1	0 / 0
Submitted	1 / 1	1 / 1	1 / 1	0 / 0
Graded	1 / 1	1 / 1	1 / 1	0 / 0
Passed hidden tests	1 / 1	0 / 1	0 / 1	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Passed hidden tests	Passed hidden tests	1 / 1	0
cli:cli/codex	Graded	Graded	0 / 1	0
mistral:codestral-latest	Graded	Graded	0 / 1	0
groq:openai/gpt-oss-120b	nothing	nothing	0 / 0	1

Stage definitions are in Appendix D. The matrix counts evaluable rollouts; lane errors are shown beside it, not inside it. The stages are not strictly nested: a model can submit without editing, and the matrix shows that rather than hiding it. Each rollout's own deepest stage is in Appendix A.

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/claude-sonnet	1	9 / 40	0	0 / 0	0	0	0	3
cli:cli/codex	1	7 / 40	0	0 / 0	0	0	0	0
mistral:codestral-latest	1	6 / 40	0	0 / 0	0	0	0	0
groq:openai/gpt-oss-120b	1	1 / 40	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/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	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

No grounding failures and no false pass claims were detected across these 4 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
cli:cli/claude-sonnet	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
cli:cli/codex	1	n/a	n/a	n/a — telemetry unavailable	0 / 1	0.0%
mistral:codestral-latest	1	40 054	40 054	not billed	0 / 1	0.0%
groq:openai/gpt-oss-120b	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

cli/claude-sonnet passed the independent hidden tests

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

1 of 3 graded rollouts completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 2 (cli/codex, codestral-latest)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted a change; the independent hidden tests still failed.

F-02 · derived from the recorded run data

F-03 LANE / INFRA

model_api_failure × 1 (openai/gpt-oss-120b)

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

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

F-03 · derived from the recorded run data

F-04 NOTE

Statistical floor — 4 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

Submitted a change; the independent hidden tests still failed.

AFFECTED LANES

cli/codex, codestral-latest

WHAT WE WOULD LOOK AT FIRST

The model completes the loop and produces a plausible near-miss — the named failing tests are the exact behavioral cases distinguishing it from the reference fix.

R–02 PROBLEM

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

AFFECTED LANES

openai/gpt-oss-120b

WHAT WE WOULD LOOK AT FIRST

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

R–03 **For anyone reading the percentages**

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

Verification

RUN ID

live-20260901T214907Z

EXAM FINGERPRINT

e8312ce98b82558f9a1c5593e7d431f409fb1024b3013b7dbd934f1398a8937f

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

f49a31336e4865c1dcefccf5141e2f6926110c72f03d6cb3d645a1761fa09ac8

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

SOURCE RECORD FILE

6bafd70a8214b1e782b25d64d4733b0b2223a2cffe518820df0a030ae5f4673b

sha256 of fr-20260901-88c3cdf1.json as written beside the standalone report.

SOURCE STANDALONE REPORT

bfb1b0dcbb7574e53573907812a76f0134b47a36ebbeb9d04ff543d189ffe58e

sha256 of fr-20260901-88c3cdf1.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

e85ac24764612e3343a739a09e26bcad398b003352b9877b03d4a5a79bdbd973

sha256 of fr-20260901-88c3cdf1.pdf.

RUN EVIDENCE BUNDLE

88c3cdf1c1b6d606e81360be7fcb06bd12c6340699acd7e3737cd37ebfce226c

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

CERTIFIED EVIDENCE SEAL

not available in this package

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
ea3e4b9b	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	93.2s
bbf9607d	cli/codex	0	submitted_failed	yes	Graded	fail	1	n/a	54.8s
441654d0	codestral-latest	0	submitted_failed	yes	Graded	fail	0	40 054	16.7s
5f24c149	openai/gpt-oss-120b	0	model_api_failure	no	Not graded — lane error	not graded	0	1 164	1129.5s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	2	Submitted a change; the independent hidden tests still failed.
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
88c3cdf1c1b6d606e81360be7fcb06bd12c6340699acd7e3737cd37ebfce226c	live-20260901T214907Z	2026-09-01T21:49:07.619617+00:00	2026-09-01T22:11:03.052288+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	a0a7dfe8cfc33cf167972072583fd76cf2fbc9fa0e3635c077b5cfcc2d907abd
rollouts/441654d0-e828-4779-91df-e35763a02c85.json	f4d362a1e4dac96705b00d459b69dd4350b888e5c02e99199e633e5eb7cc6483
rollouts/5f24c149-f696-4b36-861f-26fd7281b1bc.json	981516b5a142693648fb800d9dc7ccda9b470362261d74211e5d9afb0ff4664c
rollouts/bbf9607d-548f-4d11-9aae-777105292a62.json	9bf2824267ce3d005aa8bc66ff581cc5f433ad43bc4b241dac4218f1c19eb918
rollouts/ea3e4b9b-fe43-4913-bcbe-0c61631dd1b5.json	6c1c277aafcd3bc8378613b18967cf3d3cb7d37bdc786296494821cc430eeec2
spec	a29231f530943bca94c1b1f15da405051ea0d695923657aa869c2c5d8716de27
summary	dd09a0bb2c237da0371d9f3ef41236f0be9b6647c31303d70d6991cf19d53fd0

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 public_oss_swe): 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.

ea3e4b9b · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 371
1	list_files	toolz	ok	304
2	read_file	toolz/itertoolz.py	ok	12 166
3	read_file	toolz/itertoolz.py	ok	12 170
4	invalid_tool_call	(none)	failed	107
5	invalid_tool_call	(none)	failed	107
6	edit_file	toolz/itertoolz.py	ok	56
7	run_tests	–	ok	179
8	submit	–	ok	85

Submission diff withheld by the public-disclosure policy (58 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.

bbf9607d · cli:cli/codex · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	336
1	read_file	ISSUE.md	ok	1 371
2	read_file	toolz/itertoolz.py	ok	12 166
3	read_file	toolz/itertoolz.py	ok	12 170
4	edit_file	toolz/itertoolz.py	ok	56
5	run_tests	–	ok	179
6	submit	–	ok	85

Submission diff withheld by the public-disclosure policy (69 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.

441654d0 · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 371
1	read_file	toolz/itertoolz.py	ok	12 166
2	read_file	toolz/itertoolz.py	ok	12 170
3	read_file	toolz/itertoolz.py	ok	3 618
4	run_tests	—	ok	179
5	submit	—	ok	85

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.

5f24c149 · groq:openai/gpt-oss-120b · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	336

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 Exam · r1chardj0n3s-parse-issue-102

`public.swe.r1chardj0n3s-parse-issue-102 · record fr-20260901-4b1e3251 · record digest d94c313144f8b36d… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
4	1 of 3 graded rollouts	3 of 3 graded rollouts that called submit	1 of 4 attempts

Purpose

This report presents a controlled evaluation of 4 model lane(s) against the certified exam `public.swe.r1chardj0n3s-parse-issue-102` (v0.1.0, family public_oss_swe). Every lane received the identical frozen workspace, tool contract and limits; grading ran afterwards, in isolation, against hidden tests no lane ever saw. The purpose is to state, with reproducible evidence, what each lane did and where it stopped.

Outcome

1 of 3

graded rollouts passed the independent hidden tests and were certified — 1 of 4 attempts not graded (lane errors)

1 of 3 graded rollouts passed. cli/claude-sonnet completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. **Low-N:** 4 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was `submitted_failed` — submitted a change; the independent hidden tests still failed.

Key findings

F-01	cli/claude-sonnet passed the independent hidden tests
F-02	submitted_failed × 2 (cli/codex, codestral-latest)
F-03	model_api_failure × 1 (openai/gpt-oss-120b)
F-04	Statistical floor — 4 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: 1469.0s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the workspace, produce a correct edit, verify it with the visible tests, and submit — within 40 tool steps and 9720s 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 `public.swe.r1chardj0n3s-parse-issue-102` 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
public.swe.r1chardj0n3s-parse-issue-102	0.1.0	public_oss_swe	4 (1 per model)

Where the defect came from

Source: r1chardj0n3s/parse (MIT) · issue https://github.com/r1chardj0n3s/parse/issues/102 · upstream fix https://github.com/r1chardj0n3s/parse/pull/103

The defect and its upstream fix are public and unmodified. The exam around them — the frozen parent tree, the independently written hidden suite, the control probes and the isolation — is VVDex's, and it is what this report measures.

Certification evidence

4 of 5 certification pillars are recorded in this package, loaded from the receipts sealed into the exam and matching this run's exam fingerprint. What is recorded is shown with its real value; what is not is named below and is not claimed.

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

7 / 7

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

ATTACK PROBES

10 / 10 blocked

0 trivial exploit(s) found.

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

Certification metadata is not included in this report package for:

- sealed evidence bundle

Benchmark results remain valid as run evidence; the pillar(s) above are not substantiated by this document and this report does not claim them.

What the package does carry: the exam fingerprint below is the contract digest this run was executed against — a holder of the certification bundle for this fingerprint can join the two independently.

EXAM FINGERPRINT

483b68b383c8510607811964033 25de0e00e5f6327fd52d24eb581005be38812

The contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260901-4b1e3251
ISSUED	2026-09-01
SUBJECT EXAM	public.swe.r1chardj0n3s-parse-issue-102 · v0.1.0
FAMILY	public_oss_swe
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: partially recorded (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-01 22:11:03	Evaluation run started
2026-09-01 22:35:32	Evaluation run finished
2026-09-01 22:35:32	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
codestral-latest	mistral	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	40	9720s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	0	list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 4. Graded (evaluable) rollouts: 3. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) `cli/claude-sonnet`	1	1	1 / 1	0	0	1 / 1	255.5s	Passed hidden tests
Codex CLI (GPT, subscription) `cli/codex`	1	1	0 / 1	1	0	1 / 1	72.9s	Graded
codestral-latest `codestral-latest`	1	1	0 / 1	1	0	1 / 1	22.3s	Graded
openai/gpt-oss-120b `openai/gpt-oss-120b`	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
cli/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/codex	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
openai/gpt-oss-120b	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
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	255.5s	0.0%	none
cli/codex	0.0%	[0%, 79%]	n/a	n/a	72.9s	0.0%	submitted_failed
codestral-latest	0.0%	[0%, 79%]	73 260	not billed	22.3s	0.0%	submitted_failed
openai/gpt-oss-120b	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): cli/claude-sonnet vs cli/codex; cli/claude-sonnet vs codestral-latest; cli/claude-sonnet vs openai/gpt-oss-120b; cli/codex vs codestral-latest; cli/codex vs openai/gpt-oss-120b; codestral-latest vs openai/gpt-oss-120b. Below the 10-rollout floor for reading a rate as a rate: cli/claude-sonnet (1), cli/codex (1), codestral-latest (1), openai/gpt-oss-120b (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

The funnel is the primary reading of a run. Where a model stops says more than a single pass percentage does, especially at low N.

STAGE	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CODEX	MISTRAL:CODESTRAL-LATEST	GROQ:OPENAI/GPT-OSS-120B
Inspected	1 / 1	1 / 1	1 / 1	0 / 0
Edited	1 / 1	1 / 1	1 / 1	0 / 0
Ran tests	1 / 1	1 / 1	1 / 1	0 / 0
Submitted	1 / 1	1 / 1	1 / 1	0 / 0
Graded	1 / 1	1 / 1	1 / 1	0 / 0
Passed hidden tests	1 / 1	0 / 1	0 / 1	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Passed hidden tests	Passed hidden tests	1 / 1	0
cli:cli/codex	Graded	Graded	0 / 1	0
mistral:codestral-latest	Graded	Graded	0 / 1	0
groq:openai/gpt-oss-120b	nothing	nothing	0 / 0	1

Stage definitions are in Appendix D. The matrix counts evaluable rollouts; lane errors are shown beside it, not inside it. The stages are not strictly nested: a model can submit without editing, and the matrix shows that rather than hiding it. Each rollout's own deepest stage is in Appendix A.

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/claude-sonnet	1	11 / 40	0	0 / 0	0	0	0	3
cli:cli/codex	1	11 / 40	0	0 / 0	0	0	0	2
mistral:codestral-latest	1	9 / 40	0	0 / 0	0	0	0	1
groq:openai/gpt-oss-120b	1	1 / 40	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/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	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

No grounding failures and no false pass claims were detected across these 4 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
cli:cli/claude-sonnet	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
cli:cli/codex	1	n/a	n/a	n/a — telemetry unavailable	0 / 1	0.0%
mistral:codestral-latest	1	73 260	73 260	not billed	0 / 1	0.0%
groq:openai/gpt-oss-120b	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

cli/claude-sonnet passed the independent hidden tests

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

1 of 3 graded rollouts completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 2 (cli/codex, codestral-latest)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted a change; the independent hidden tests still failed.

F-02 · derived from the recorded run data

F-03 LANE / INFRA

model_api_failure × 1 (openai/gpt-oss-120b)

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

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

F-03 · derived from the recorded run data

F-04 NOTE

Statistical floor — 4 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

Submitted a change; the independent hidden tests still failed.

AFFECTED LANES

cli/codex, codestral-latest

WHAT WE WOULD LOOK AT FIRST

The model completes the loop and produces a plausible near-miss — the named failing tests are the exact behavioral cases distinguishing it from the reference fix.

R–02 PROBLEM

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

AFFECTED LANES

openai/gpt-oss-120b

WHAT WE WOULD LOOK AT FIRST

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

R–03 **For anyone reading the percentages**

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

Verification

RUN ID

live-20260901T221103Z

EXAM FINGERPRINT

483b68b383c851060781196403325de0e00e5f6327fd52d24eb581005be38812

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

d94c313144f8b36db4448440b873462bc9372e1783b1164451695b7b8497816a

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

SOURCE RECORD FILE

31dd2cfcaf563ac9ae7c31d50b5d8adf9d8ebbc97646767d922b02731ad62f88

sha256 of fr-20260901-4b1e3251.json as written beside the standalone report.

SOURCE STANDALONE REPORT

3410a4fc424b582f2fd03119f69ff337cd9785ec620503730755c9b78379987b

sha256 of fr-20260901-4b1e3251.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

bddb2882e46b9be29bfdffe30188294705be407c601284a7b2caef6c7d697155

sha256 of fr-20260901-4b1e3251.pdf.

RUN EVIDENCE BUNDLE

4b1e3251d3e5a146a99f2e348bb71c76953245e7f85d35c8f5a1a24011faa5ed

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

CERTIFIED EVIDENCE SEAL

not available in this package

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
14c94f8c	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	3	n/a	255.5s
1f2eb2de	cli/codex	0	submitted_failed	yes	Graded	fail	3	n/a	72.9s
890eafa7	codestral-latest	0	submitted_failed	yes	Graded	fail	1	73 260	22.3s
c4fb19d4	openai/gpt-oss-120b	0	model_api_failure	no	Not graded — lane error	not graded	0	1 381	1097.0s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	2	Submitted a change; the independent hidden tests still failed.
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
4b1e3251d3e5a146a99f2e348bb71c76953245e7f85d35c8f5a1a24011faa5ed	live-20260901T221103Z	2026-09-01T22:11:03.440919+00:00	2026-09-01T22:35:32.433884+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	d782d1af3bde0176db198da9b48bc877239d6988eaf07e48540e16fd4be4bbce
rollouts/14c94f8c-9a21-4fab-9d6d-4b3d379e5fa1.json	be97be4950dccc8cf96c33573f8b3ce5dc5387afcac804f422d6c0ac137c4b20
rollouts/1f2eb2de-ddf7-4ca8-88b9-392f2a829f01.json	881c103264c5bc44358ab0fcb4229d0f8387f97d453c16ac2c762250e045189f
rollouts/890eafa7-f9d7-4805-818b-8003d1157013.json	4cb38d26a39c49f2799928f1f2209d5a1f31ea71f402dc84b35e832dcfc943bd
rollouts/c4fb19d4-b5b5-4cc2-a6cf-0c9e2c815b04.json	b77ccd2450443c6339c9e99d4ec89af7e9bd83a7e2edaca195967dfdac9f03c4
spec	b15c867013aeb4603011ad5b63f4244fa12213d3a3bc8e3cb944efebb47b2d81
summary	1a300c4fdd48b1dedc2b7632e5a495d39413e29f668ade037463d474421b7273

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 public_oss_swe): 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.

14c94f8c · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 181
1	read_file	parse.py	ok	12 166
2	read_file	parse.py	ok	12 170
3	read_file	parse.py	ok	12 170
4	read_file	parse.py	ok	12 170
5	invalid_tool_call	(none)	failed	107
6	edit_file	parse.py	ok	46
7	edit_file	parse.py	ok	46
8	edit_file	parse.py	ok	46
9	run_tests	—	ok	179
10	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (101 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.

1f2eb2de · cli:cli/codex · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	36
1	read_file	ISSUE.md	ok	1 181
2	read_file	parse.py	ok	12 166
3	read_file	parse.py	ok	12 170
4	read_file	parse.py	ok	12 170
5	read_file	parse.py	ok	12 170
6	edit_file	parse.py	ok	46
7	edit_file	parse.py	ok	46
8	edit_file	parse.py	ok	46
9	run_tests	—	ok	179

STEP	TOOL	PATH	RESULT	CHARS
10	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (74 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.

890eafa7 · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	36
1	read_file	parse.py	ok	12 166
2	read_file	parse.py	ok	12 170
3	read_file	parse.py	ok	12 170
4	read_file	parse.py	ok	924
5	invalid_tool_call	(none)	failed	107
6	edit_file	parse.py	ok	46
7	run_tests	—	ok	179
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (12 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.

c4fb19d4 · groq:openai/gpt-oss-120b · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	36

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 Exam · grounded-rag-1

`vvdex.knowledge.grounded-rag-1 · record fr-20260901-718db345 · record digest a6426b0b7932b6c0… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
4	1 of 3 graded rollouts	3 of 3 graded rollouts that called submit	1 of 4 attempts

Purpose

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

Outcome

1 of 3

graded rollouts passed the independent hidden tests and were certified — 1 of 4 attempts not graded (lane errors)

1 of 3 graded rollouts passed. cli/codex completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. **Low-N:** 4 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was `submitted_failed` — submitted a change; the independent hidden tests still failed.

Key findings

F-01	cli/codex passed the independent hidden tests
F-02	submitted_failed × 2 (cli/claude-sonnet, codestral-latest)
F-03	model_api_failure × 1 (openai/gpt-oss-120b)
F-04	Statistical floor — 4 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: 1136.8s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the workspace, produce a correct edit, verify it with the visible tests, and submit — within 30 tool steps and 7320s 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.knowledge.grounded-rag-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.knowledge.grounded-rag-1	0.1.0	rag_grounding	4 (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

4 of 5 certification pillars are recorded in this package, loaded from the receipts sealed into the exam and matching this run's exam fingerprint. What is recorded is shown with its real value; what is not is named below and is not claimed.

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

7 / 7

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

ATTACK PROBES

10 / 10 blocked

0 trivial exploit(s) found.

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

Certification metadata is not included in this report package for:

- sealed evidence bundle

Benchmark results remain valid as run evidence; the pillar(s) above are not substantiated by this document and this report does not claim them.

What the package does carry: the exam fingerprint below is the contract digest this run was executed against — a holder of the certification bundle for this fingerprint can join the two independently.

EXAM FINGERPRINT

4880a527ef97ca2f3fe6b57be026c3dc1be51e9e104d79fc13a3e558331dda92

The contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260901-718db345
ISSUED	2026-09-01
SUBJECT EXAM	vvdex.knowledge.grounded-rag-1 · v0.1.0
FAMILY	rag_grounding
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: partially recorded (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-01 21:30:10	Evaluation run started
2026-09-01 21:49:07	Evaluation run finished
2026-09-01 21:49:07	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
codestral-latest	mistral	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	30	7320s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	40	list_files, read_file, write_file, edit_file, run_tests, submit, run_command

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: 4. Graded (evaluable) rollouts: 3. Not graded — lane errors: 1. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) `cli/claude-sonnet`	1	1	0 / 1	1	0	1 / 1	71.8s	Graded
Codex CLI (GPT, subscription) `cli/codex`	1	1	1 / 1	0	0	1 / 1	68.5s	Passed hidden tests
codestral-latest `codestral-latest`	1	1	0 / 1	1	0	1 / 1	18.7s	Graded
openai/gpt-oss-120b `openai/gpt-oss-120b`	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
cli/claude-sonnet	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
cli/codex	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
openai/gpt-oss-120b	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
cli/claude-sonnet	0.0%	[0%, 79%]	n/a	n/a	71.8s	0.0%	submitted_failed
cli/codex	100.0%	[21%, 100%]	n/a	n/a	68.5s	0.0%	none
codestral-latest	0.0%	[0%, 79%]	27 288	not billed	18.7s	0.0%	submitted_failed
openai/gpt-oss-120b	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): cli/claude-sonnet vs cli/codex; cli/claude-sonnet vs codestral-latest; cli/claude-sonnet vs openai/gpt-oss-120b; cli/codex vs codestral-latest; cli/codex vs openai/gpt-oss-120b; codestral-latest vs openai/gpt-oss-120b. Below the 10-rollout floor for reading a rate as a rate: cli/claude-sonnet (1), cli/codex (1), codestral-latest (1), openai/gpt-oss-120b (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

The funnel is the primary reading of a run. Where a model stops says more than a single pass percentage does, especially at low N.

STAGE	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CODEX	MISTRAL:CODESTRAL-LATEST	GROQ:OPENAI/GPT-OSS-120B
Inspected	1 / 1	1 / 1	1 / 1	0 / 0
Edited	1 / 1	1 / 1	1 / 1	0 / 0
Ran tests	1 / 1	1 / 1	1 / 1	0 / 0
Submitted	1 / 1	1 / 1	1 / 1	0 / 0
Graded	1 / 1	1 / 1	1 / 1	0 / 0
Passed hidden tests	0 / 1	1 / 1	0 / 1	0 / 0

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Graded	Graded	0 / 1	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	1 / 1	0
mistral:codestral-latest	Graded	Graded	0 / 1	0
groq:openai/gpt-oss-120b	nothing	nothing	0 / 0	1

Stage definitions are in Appendix D. The matrix counts evaluable rollouts; lane errors are shown beside it, not inside it. The stages are not strictly nested: a model can submit without editing, and the matrix shows that rather than hiding it. Each rollout's own deepest stage is in Appendix A.

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/claude-sonnet	1	12 / 30	0	0 / 40	0	0	0	0
cli:cli/codex	1	14 / 30	0	0 / 40	0	0	0	0
mistral:codestral-latest	1	10 / 30	0	0 / 40	0	0	0	1
groq:openai/gpt-oss-120b	1	4 / 30	0	0 / 40	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/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	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

No grounding failures and no false pass claims were detected across these 4 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
cli:cli/claude-sonnet	1	n/a	n/a	n/a — telemetry unavailable	0 / 1	0.0%
cli:cli/codex	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
mistral:codestral-latest	1	27 288	27 288	not billed	0 / 1	0.0%
groq:openai/gpt-oss-120b	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

cli/codex passed the independent hidden tests

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

1 of 3 graded rollouts completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 2 (cli/claude-sonnet, codestral-latest)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted a change; the independent hidden tests still failed.

F-02 · derived from the recorded run data

F-03 LANE / INFRA

model_api_failure × 1 (openai/gpt-oss-120b)

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

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

F-03 · derived from the recorded run data

F-04 NOTE

Statistical floor — 4 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

Submitted a change; the independent hidden tests still failed.

AFFECTED LANES

cli/claude-sonnet, codestral-latest

WHAT WE WOULD LOOK AT FIRST

The model completes the loop and produces a plausible near-miss — the named failing tests are the exact behavioral cases distinguishing it from the reference fix.

R–02 PROBLEM

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

AFFECTED LANES

openai/gpt-oss-120b

WHAT WE WOULD LOOK AT FIRST

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

R–03 **For anyone reading the percentages**

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

Verification

RUN ID

live-20260901T213010Z

EXAM FINGERPRINT

4880a527ef97ca2f3fe6b57be026c3dc1be51e9e104d79fc13a3e558331dda92

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

a6426b0b7932b6c0f679f230c44c4b75a30df10a17f7e145b567ac9849d63a36

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

SOURCE RECORD FILE

1d818470e53ea53ee6497e9038e19517a4d827d382d3f335018284b31c0e8002

sha256 of fr-20260901-718db345.json as written beside the standalone report.

SOURCE STANDALONE REPORT

b287b1728db1ec9c26f6c714a3a7afa404faa12832740c8a9cbd56d0807a9c36

sha256 of fr-20260901-718db345.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

079770731b29ce711d1b3ea3cef453fcb13807ccec7318f06647d5b85c0e12de

sha256 of fr-20260901-718db345.pdf.

RUN EVIDENCE BUNDLE

718db345e9c729728f9b9272bf0b2ede5774a680273099943b6a7d7bdfce9108

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

CERTIFIED EVIDENCE SEAL

not available in this package

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
a8b6ca0b	cli/claude-sonnet	0	submitted_failed	yes	Graded	fail	1	n/a	71.8s
f9d23573	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	68.5s
2c0b85d4	codestral-latest	0	submitted_failed	yes	Graded	fail	1	27 288	18.7s
97723e61	openai/gpt-oss-120b	0	model_api_failure	no	Not graded — lane error	not graded	0	7 583	956.6s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	2	Submitted a change; the independent hidden tests still failed.
model_api_failure	1	The model endpoint failed to answer. Not a verdict about the model's ability.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
718db345e9c729728f9b9272bf0b2ede5774a68027309994 3b6a7d7bdfce9108	live-20260901T213010Z	2026-09-01T21:30:10.375515+00:00	2026-09-01T21:49:07.174353+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	0cf64390cffbfc2dd80444c3bfda61950565191526629d2c3793739421d4d2c6
rollouts/2c0b85d4-ba0d-4c0c-8216-ea243bfbaab9.json	0abbbfe6f3b34f1eb205a6a3b3dbfbe512347c36a6af75aed38d37a69d8c0218
rollouts/97723e61-d56d-4aed-8b61-dee14d643db8.json	efcf0a5acc2b90fe8ec1305a21e14b08ddeee9c076a44e03bf6bad6f10dacbd3
rollouts/a8b6ca0b-1ed7-4a52-a051-1055108b710b.json	3dda83791f608a375d207c90d6d5a078feb5723d64da05d46f42ff0a518cab6a
rollouts/f9d23573-24ee-4f45-89e9-2771b0bafc68.json	890b8ef9a1178d3c7ecb764f5b06a28aa57a5d3e509c3ed2d3d62e5397571638
spec	0b760c0ff8444d7143b455fd1603165f7f4adfc0dec6f1c792d573624333536f
summary	19ec6555d12e20cc823c18b06ae11821257eac563b184d022824e4198d81141d

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 — engine registry; family rag_grounding): 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.

a8b6ca0b · cli:cli/claude-sonnet · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	read_file	corpus/data-protection-policy-v3.txt	ok	293
3	read_file	corpus/retention-addendum-2026.txt	ok	142
4	read_file	corpus/org-chart-2026.txt	ok	183
5	read_file	corpus/incident-report-2026.txt	ok	169
6	read_file	corpus/security-whitepaper.txt	ok	142
7	read_file	corpus/employee-handbook.txt	ok	167
8	read_file	corpus/vendor-list.txt	ok	153
9	write_file	answers/answers.json	ok	69
10	run_tests	—	ok	181
11	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (45 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.

f9d23573 · cli:cli/codex · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	list_files	.	ok	287
2	read_file	questions.json	ok	883
3	read_file	corpus/data-protection-policy-v1.txt	ok	239
4	read_file	corpus/data-protection-policy-v3.txt	ok	293
5	read_file	corpus/employee-handbook.txt	ok	167
6	read_file	corpus/incident-report-2026.txt	ok	169
7	read_file	corpus/org-chart-2026.txt	ok	183
8	read_file	corpus/retention-addendum-2026.txt	ok	142

STEP	TOOL	PATH	RESULT	CHARS
9	read_file	corpus/security-whitepaper.txt	ok	142
10	read_file	corpus/vendor-list.txt	ok	153
11	write_file	answers/answers.json	ok	69
12	run_tests	–	ok	181
13	submit	–	ok	85

Submission diff withheld by the public-disclosure policy (45 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.

2c0b85d4 · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	read_file	–	failed	35
3	read_file	–	failed	35
4	list_files	corpus	ok	249
5	read_file	corpus/data-protection-policy-v3.txt	ok	293
6	read_file	corpus/retention-addendum-2026.txt	ok	142
7	write_file	answers/answers.json	ok	69
8	run_tests	–	ok	181
9	submit	–	ok	85

Submission diff withheld by the public-disclosure policy (45 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.

97723e61 · groq:openai/gpt-oss-120b · seed 0 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	ISSUE.md	ok	1 953
2	list_files	.	ok	287
3	read_file	questions.json	ok	883

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.

13 Exam · memory-fact-update-1

`vvdex.knowledge.memory-fact-update-1 · record fr-20260901-109cfa3c · record digest 6354cff6fc47eca1… · harness clean · containment clean`

Executive summary

MODELS TESTED	CERTIFIED PASSES	SUBMITTED	NOT GRADED — LANE ERRORS
4	3 of 4 graded rollouts	4 of 4 graded rollouts that called submit	0 of 4 attempts

Purpose

This report presents a controlled evaluation of 4 model lane(s) against the certified exam `vvdex.knowledge.memory-fact-update-1` (v0.1.0, family long_term_memory). 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 4 graded rollouts passed the independent hidden tests and were certified

3 of 4 graded rollouts passed. cli/claude-sonnet, cli/codex, openai/gpt-oss-120b completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. **Low-N:** 4 rollout(s) is below the 10-rollout floor for reading a rate as a rate — read the per-rollout outcomes, not the percentages. The most common way to fail was `submitted_failed` — submitted a change; the independent hidden tests still failed.

Key findings

F-01	cli/claude-sonnet, cli/codex, openai/gpt-oss-120b passed the independent hidden tests
F-02	submitted_failed × 1 (codestral-latest)
F-03	Statistical floor — 4 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: 122.3s.

Exam definition

In scope. The ability of each configured model lane to complete this one certified task end to end — inspect the workspace, produce a correct edit, verify it with the visible tests, and submit — within 20 tool steps and 4920s 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.knowledge.memory-fact-update-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.knowledge.memory-fact-update-1	0.1.0	long_term_memory	4 (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

4 of 5 certification pillars are recorded in this package, loaded from the receipts sealed into the exam and matching this run's exam fingerprint. What is recorded is shown with its real value; what is not is named below and is not claimed.

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

7 / 7

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

ATTACK PROBES

10 / 10 blocked

0 trivial exploit(s) found.

RUNTIME

sha256:679b36225a1f83238efba8c71b9cde3c24caaeacc9b682ae92819cb55eec328f

network policy: deny

Certification metadata is not included in this report package for:

- sealed evidence bundle

Benchmark results remain valid as run evidence; the pillar(s) above are not substantiated by this document and this report does not claim them.

What the package does carry: the exam fingerprint below is the contract digest this run was executed against — a holder of the certification bundle for this fingerprint can join the two independently.

EXAM FINGERPRINT

2052eec0ff2d3044dc269e3599ae1bfaa45d0d8cd5df9554796105538d226c21

The contract digest this run was executed against.

Evaluation configuration

REPORT	fr-20260901-109cfa3c
ISSUED	2026-09-01
SUBJECT EXAM	vvdex.knowledge.memory-fact-update-1 · v0.1.0
FAMILY	long_term_memory
PREPARED BY	VVDex Forge engine vvdex-env 0.1.0
CLASSIFICATION	Independent model evaluation report — independently verifiable
CERTIFICATION	Exam certification: partially recorded (see Certification evidence)
STATUS	FINAL · report sealed against its own digest

Timeline

WHEN (UTC)	EVENT
2026-09-01 21:28:07	Evaluation run started
2026-09-01 21:30:10	Evaluation run finished
2026-09-01 21:30:10	Report generated from the sealed run evidence

Models under test

MODEL	PROVIDER	ENDPOINT	BILLING
Claude Code CLI (Sonnet, subscription)	cli	local runtime	operator subscription, flat — not billed per token
Codex CLI (GPT, subscription)	cli	local runtime	operator subscription, flat — not billed per token
codestral-latest	mistral	VVDex-configured	VVDex free-quota
openai/gpt-oss-120b	groq	VVDex-configured	VVDex free-quota

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

Limits

RUNTIME	STEP LIMIT	WALL-CLOCK LIMIT	MAX OUTPUT TOKENS
docker	20	4920s	2048

TEMPERATURE	RETRY POLICY	COMMAND BUDGET	TOOLS OFFERED
0.2	none	24	list_files, read_file, write_file, edit_file, run_tests, submit, run_command

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: 4. Graded (evaluable) rollouts: 4. 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
Claude Code CLI (Sonnet, subscription) `cli/claude-sonnet`	1	1	1 / 1	0	0	1 / 1	18.2s	Passed hidden tests
Codex CLI (GPT, subscription) `cli/codex`	1	1	1 / 1	0	0	1 / 1	27.8s	Passed hidden tests
codestral-latest `codestral-latest`	1	1	0 / 1	1	0	1 / 1	6.8s	Graded
openai/gpt-oss-120b `openai/gpt-oss-120b`	1	1	1 / 1	0	0	1 / 1	48.4s	Passed hidden tests

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/claude-sonnet	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
cli/codex	1	1	0	1	100.0%	[21%, 100%]	0	100.0%
codestral-latest	1	1	0	0	0.0%	[0%, 79%]	0	0.0%
openai/gpt-oss-120b	1	1	0	1	100.0%	[21%, 100%]	0	100.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/claude-sonnet	100.0%	[21%, 100%]	n/a	n/a	18.2s	0.0%	none
cli/codex	100.0%	[21%, 100%]	n/a	n/a	27.8s	0.0%	none
codestral-latest	0.0%	[0%, 79%]	5 785	not billed	6.8s	0.0%	submitted_failed
openai/gpt-oss-120b	100.0%	[21%, 100%]	7 980	not billed	48.4s	0.0%	none

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

This is not a leaderboard. Not separated (intervals overlap): cli/claude-sonnet vs cli/codex; cli/claude-sonnet vs codestral-latest; cli/claude-sonnet vs openai/gpt-oss-120b; cli/codex vs codestral-latest; cli/codex vs openai/gpt-oss-120b; codestral-latest vs openai/gpt-oss-120b. Below the 10-rollout floor for reading a rate as a rate: cli/claude-sonnet (1), cli/codex (1), codestral-latest (1), openai/gpt-oss-120b (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

The funnel is the primary reading of a run. Where a model stops says more than a single pass percentage does, especially at low N.

STAGE	CLI:CLI/CLAUDE-SONNET	CLI:CLI/CODEX	MISTRAL:CODESTRAL-LATEST	GROQ:OPENAI/GPT-OSS-120B
Inspected	1 / 1	1 / 1	1 / 1	1 / 1
Edited	1 / 1	1 / 1	1 / 1	1 / 1
Ran tests	1 / 1	1 / 1	1 / 1	1 / 1
Submitted	1 / 1	1 / 1	1 / 1	1 / 1
Graded	1 / 1	1 / 1	1 / 1	1 / 1
Passed hidden tests	1 / 1	1 / 1	0 / 1	1 / 1

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Passed hidden tests	Passed hidden tests	1 / 1	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	1 / 1	0
mistral:codestral-latest	Graded	Graded	0 / 1	0
groq:openai/gpt-oss-120b	Passed hidden tests	Passed hidden tests	1 / 1	0

Stage definitions are in Appendix D. The matrix counts evaluable rollouts; lane errors are shown beside it, not inside it. The stages are not strictly nested: a model can submit without editing, and the matrix shows that rather than hiding it. Each rollout's own deepest stage is in Appendix A.

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/claude-sonnet	1	5 / 20	0	0 / 24	0	0	0	0
cli:cli/codex	1	6 / 20	0	0 / 24	0	0	0	0
mistral:codestral-latest	1	4 / 20	0	0 / 24	0	0	0	0
groq:openai/gpt-oss-120b	1	6 / 20	0	0 / 24	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.

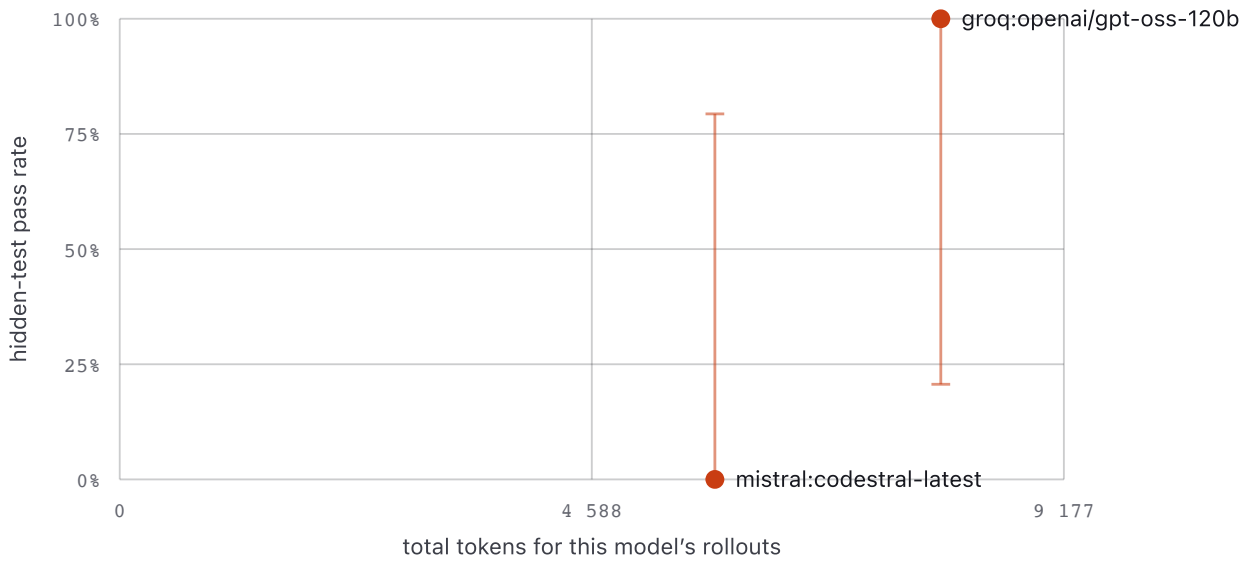
MODEL	N	GROUNDING FAILURES	KINDS	ROLLOUTS AFFECTED	OVERCLAIMS	OVERCLAIM RATE	RE-READS
cli:cli/claude-sonnet	1	0	none	0 / 1	0 / 1	0.0%	0
cli:cli/codex	1	0	none	0 / 1	0 / 1	0.0%	0
mistral:codestral-latest	1	0	none	0 / 1	0 / 1	0.0%	0
groq:openai/gpt-oss-120b	1	0	none	0 / 1	0 / 1	0.0%	0

No grounding failures and no false pass claims were detected across these 4 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
cli:cli/claude-sonnet	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
cli:cli/codex	1	n/a	n/a	n/a — telemetry unavailable	1 / 1	100.0%
mistral:codestral-latest	1	5 785	5 785	not billed	0 / 1	0.0%
groq:openai/gpt-oss-120b	1	7 980	7 980	not billed	1 / 1	100.0%

Findings

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

F-01 PASS

cli/claude-sonnet, cli/codex, openai/gpt-oss-120b passed the independent hidden tests

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

3 of 4 graded rollouts completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. This demonstrates the task is solvable under the published limits.

F-01 · derived from the recorded run data

F-02 ATTENTION

submitted_failed × 1 (codestral-latest)

Severity: ATTENTION | Evidence: Appendix A · Appendix C

Submitted a change; the independent hidden tests still failed.

F-02 · derived from the recorded run data

F-03 NOTE

Statistical floor — 4 rollout(s) is below the 10-rollout rate floor

Severity: NOTE | Evidence: Appendix D

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

F-03 · derived from the recorded run data

Recommendations

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

R–01

PROBLEM

Submitted a change; the independent hidden tests still failed.

AFFECTED LANES

codestral-latest

WHAT WE WOULD LOOK AT FIRST

The model completes the loop and produces a plausible near-miss — the named failing tests are the exact behavioral cases distinguishing it from the reference fix.

R–02

For anyone reading the percentages

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

Verification

RUN ID

live-20260901T212807Z

EXAM FINGERPRINT

2052eec0ff2d3044dc269e3599ae1bfaa45d0d8cd5df9554796105538d226c21

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

6354cff6fc47eca13b596d285f805a9f3994d048d69b446166acfeb7f8659929

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

SOURCE RECORD FILE

57be625c483af8a4fd88cfae9aaa8a5e6f8678cd89edf7fcc6d4bce490c5f405

sha256 of fr-20260901-109cfa3c.json as written beside the standalone report.

SOURCE STANDALONE REPORT

9c2db3fff15cc6edb9ee1cc71c15a45911a5fdc7e9b88984220898a5f71eb12f

sha256 of fr-20260901-109cfa3c.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

1419c72eff2ac1b20c29b732009b25783591a2234a6f5281c6bbfcbcace85b77

sha256 of fr-20260901-109cfa3c.pdf.

RUN EVIDENCE BUNDLE

109cfa3cb6a27c4876a0b1e77b7655db20cbf3d7d5a4bc821f9b5544568c5ff7

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

CERTIFIED EVIDENCE SEAL

not available in this package

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
267a40b8	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	18.2s
59ba63e4	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	27.8s
468a0f6c	codestral-latest	0	submitted_failed	yes	Graded	fail	1	5 785	6.8s
a2aa98b9	openai/gpt-oss-120b	0	submitted_passed	yes	Passed hidden tests	pass	1	7 980	48.4s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	1	Submitted a change; the independent hidden tests still failed.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST	RUN ID	STARTED	FINISHED
109cfa3cb6a27c4876a0b1e77b7655db20cbf3d7d5a4bc821f9b5544568c5ff7	live-20260901T212807Z	2026-09-01T21:28:07.881781+00:00	2026-09-01T21:30:10.198797+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	21cb3b6801e9ba0fbfde7aef29d35f722640e917b4f2e55f25a831f886216304
rollouts/267a40b8-e381-4e92-a3d0-98520dac5c21.json	c288894288ae41ccdcda9328e61981581917d3e3b7825460e473e745b00e6698
rollouts/468a0f6c-89d2-4e79-a529-9e65f466d232.json	a665188e839b85232fe95320387767209c135ed09ab9a1e47919722e92f737e2
rollouts/59ba63e4-1d43-44e2-9e3a-9de7ee90b848.json	59ac9e2ec584ad5b69bc9268bf9c764416c8e960ea22da3541b721ca8f089f38
rollouts/a2aa98b9-d5c5-4971-8689-b4f8561ac5ff.json	5720d045e7ff25175a10ade657e521b9e648190596a51ec0ff529f173dffa5ac
spec	672ae47bc67fb407b31949f6e69d215424e4d6d2c03ec838a55924b88776d0e3
summary	d424d32f949d89fca759517f140b1af38e709799822fab8d571768b74c6b93ac

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 — engine registry; family long_term_memory): 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.

267a40b8 · cli:cli/claude-sonnet · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	memory/notes.md	ok	135
1	read_file	facts/current.json	ok	65
2	write_file	answers/answer.json	ok	68
3	run_tests	–	ok	181
4	submit	–	ok	85

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

59ba63e4 · cli:cli/codex · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	read_file	facts/current.json	ok	65
2	read_file	memory/notes.md	ok	135
3	write_file	answers/answer.json	ok	68
4	run_tests	–	ok	181
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.

468a0f6c · mistral:codestral-latest · seed 0 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	memory/notes.md	ok	135
1	write_file	answers/answer.json	ok	68
2	run_tests	–	ok	181
3	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.

a2aa98b9 · groq:openai/gpt-oss-120b · seed 0 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	`.`	ok	57
1	read_file	`facts/current.json`	ok	65
2	read_file	`memory/notes.md`	ok	135
3	write_file	`answers/answer.json`	ok	68
4	run_tests	—	ok	181
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.

14 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`.

- public.swe.martinblech-xmltodict-issue-257 → fr-20260901-a27316a3 · recordDigest 96f8b631c6d65f03b21e4a9a94e86c82874e1506fac38c9a093d2f8c618dd2dc
- public.swe.pytoolz-toolz-issue-602 → fr-20260901-88c3cdf1 · recordDigest f49a31336e4865c1dcefccf5141e2f6926110c72f03d6cb3d645a1761fa09ac8
- public.swe.r1chardj0n3s-parse-issue-102 → fr-20260901-4b1e3251 · recordDigest d94c313144f8b36db4448440b873462bc9372e1783b1164451695b7b8497816a
- vvdex.knowledge.grounded-rag-1 → fr-20260901-718db345 · recordDigest a6426b0b7932b6c0f679f230c44c4b75a30df10a17f7e145b567ac9849d63a36
- vvdex.knowledge.memory-fact-update-1 → fr-20260901-109cfa3c · recordDigest 6354cff6fc47eca13b596d285f805a9f3994d048d69b446166acfeb7f8659929

15 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.