

Featured exams — ten rollouts per lane

3 sealed evaluations · 89 graded rollouts · 52
certified passes

HISTORICAL — SUPERSEDED
CAMPAIGN REPORT

2026-09-02

Evaluation scope:
cli/claude-sonnet
cli/codex
codestral-latest

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

fc-3194f803055c · 3 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	4
03.	Environment families	5
04.	Campaign aggregate across these three exams	6
05.	Harness status	7
06.	Cells excluded from capability	8
07.	What the outcomes mean	9
08.	What this report withholds, and why	10
09.	Exam · martinblech-xmltodict-issue-257	11
10.	Exam · grounded-rag-1	47
11.	Exam · memory-fact-update-1	75
12.	Evidence	98
13.	Attestation	99

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Release status: HISTORICAL — superseded.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 3	SEALED RECORDS 3	MODEL LANES 3	ATTEMPTS 90
GRADED ROLLOUTS 89	CERTIFIED PASSES 52	LANE ERRORS (NOT GRADED) 1	WITHHELD / INVALID 0

Purpose

This report presents a controlled evaluation of 3 model lane(s) against 3 fixed certified exams, 10 rollouts 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

52

certified pass(es) across 89 graded rollouts (90 attempts, 1 lane error(s) not graded) in 9 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-3194f803055c · generated from 3 verified sealed record(s)

Key findings across the campaign

- martinblech-xmltodict-issue-257 · cli/claude-sonnet · **ATTENTION** — 9 certified passes / 10 graded rollouts; 1 model failure; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- martinblech-xmltodict-issue-257 · cli/codex · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- martinblech-xmltodict-issue-257 · codestral-latest · **ATTENTION** — 0 certified passes / 10 graded rollouts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- grounded-rag-1 · cli/claude-sonnet · **ATTENTION** — 1 certified passes / 10 graded rollouts; 9 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- grounded-rag-1 · cli/codex · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- grounded-rag-1 · codestral-latest · **ATTENTION** — 0 certified passes / 9 graded rollouts; 9 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).
- memory-fact-update-1 · cli/claude-sonnet · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- memory-fact-update-1 · cli/codex · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- memory-fact-update-1 · codestral-latest · **ATTENTION** — 2 certified passes / 10 graded rollouts; 8 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	SWE PUBLIC.SWE.MARTINBLECH-XMLTODI CT-ISSUE-257	RAG VVDEX.KNOWLEDGE.GROUNDED -RAG-1	MEMORY VVDEX.KNOWLEDGE.MEMORY-FACT -UPDATE-1
cli/codex	10/10	10/10	10/10
cli/claude-sonnet	9/10	1/10	10/10
codestral-latest	0/10	0/9 + 1 lane error	2/10

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 three 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.
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 three exams

MODEL LANE	CERTIFIED PASSES	GRADED ROLLOUTS	LANE ERRORS	WITHHELD	INVALID	CERTIFIED RATE (95% CI)
cli/codex	30	30	0	0	0	100% [89-100]
cli/claude-sonnet	20	30	0	0	0	67% [49-81]
codestral-latest	2	29	1	0	0	7% [2-22]

Raw counts first. *Graded rollouts* = certified passes + model fails; lane errors, withheld and invalid cells are outside capability arithmetic in both directions. A rate with its 95% Wilson interval is shown only for a lane with at least 10 countable attempts; intervals that overlap support no ordering. Every percentage in this section is an aggregate over this campaign's exams only, not a capability score or a ranking.

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
grounded-rag-1	codestral-latest	4	lane error	HTTP 400: Assistant message must have either content or tool_calls, but not none.

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
<code>martinblech-xmltodict-issue-257</code>	<code>public_source</code>	Applied consistently to 1 sealed records. public-source disclosure (ownership <code>public_source</code> declared by engine registry, upstream https://github.com/martinblech/xmltodict under MIT; family <code>public_oss_swe</code>): the submitted change is a patch against a public repository, so a redacted excerpt of it is published.
<code>grounded-rag-1</code>	<code>answer_key</code>	Applied consistently to 1 sealed records. answer-key disclosure (ownership <code>vvdex_proprietary</code> — engine registry; family <code>rag_grounding</code>): 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.
<code>memory-fact-update-1</code>	<code>answer_key</code>	Applied consistently to 1 sealed records. answer-key disclosure (ownership <code>vvdex_proprietary</code> — engine registry; family <code>long_term_memory</code>): 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-20260902-6616b188 · record digest 138425da59aabc18... · harness c
lean · containment clean

Executive summary

MODELS TESTED 3	CERTIFIED PASSES 19 of 30 graded rollouts	SUBMITTED 30 of 30 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 30 attempts
---------------------------------------	--	--	--

Purpose

This report presents a controlled evaluation of 3 model lane(s) against the certified exam pu
blic.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

19 of 30

graded rollouts passed the independent hidden tests and were certified

19 of 30 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.
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 passed the independent hidden tests

F-02 submitted_failed × 11 (cli/claude-sonnet, codestral-latest)

Each finding is stated in full, with evidence, in the Findings section; recommendations for
acting on them follow it. Elapsed wall clock for the whole engagement: 2327.3s.

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 40 tool steps and 9720s of wall clock, at 10 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

The instrument. Exam `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	30 (10 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-20260902-6616b188
ISSUED	2026-09-02
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
2026-09-02 00:18:47	Evaluation run started
2026-09-02 00:57:34	Evaluation run finished
2026-09-02 00:57:34	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

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

Limits

RUNTIME docker	STEP LIMIT 40	WALL-CLOCK LIMIT 9720s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 0	TOOLS OFFERED list_files, read_file, write_file, edit_file, run_tests, submit

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

Model outcomes

Denominators. Attempts requested: 30. Graded (evaluable) rollouts: 30. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	10	10	9 / 10	1	0	10 / 10	159.8s	Passed hidden tests
Codex CLI (GPT, subscription) cli/codex	10	10	10 / 10	0	0	10 / 10	42.4s	Passed hidden tests
codestral-latest codestral-latest	10	10	0 / 10	10	0	10 / 10	11.5s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
cli/claude-sonnet	10	10	0	9	90.0%	[60%, 98%]	0–9	90.0%	100.0%	100.0%	100.0%
cli/codex	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
codestral-latest	10	10	0	0	0.0%	[0%, 28%]	0–9	0.0%	0.0%	0.0%	0.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/claude-sonnet	90.0%	[60%, 98%]	n/a	n/a	159.8s	0.0%	submitted_failed
cli/codex	100.0%	[72%, 100%]	n/a	n/a	42.4s	0.0%	none
codestral-latest	0.0%	[0%, 28%]	172 890	not billed	11.5s	0.0%	submitted_failed

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

This is not a leaderboard. Separated by this run (intervals do not overlap): cli/claude-sonnet vs codestral-latest; cli/codex vs codestral-latest. Not separated (intervals overlap): cli/claude-sonnet vs cli/codex. 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
Inspected	10 / 10	10 / 10	10 / 10
Edited	9 / 10	10 / 10	10 / 10
Ran tests	10 / 10	10 / 10	10 / 10
Submitted	10 / 10	10 / 10	10 / 10
Graded	10 / 10	10 / 10	10 / 10
Passed hidden tests	9 / 10	10 / 10	0 / 10

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	9 / 10	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	10 / 10	0
mistral:codestral-latest	Graded	Graded	0 / 10	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	10	5 / 40	0	0 / 0	0	0	0	0
cli:cli/codex	10	6 / 40	0	0 / 0	0	0	0	0
mistral:codestral-latest	10	5 / 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	10	0	none	0 / 10	0 / 10	0.0%	0
cli:cli/codex	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:codestral-latest	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 30 rollout(s). 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	10	n/a	n/a	n/a – telemetry unavailable	9 / 10	90.0%
cli:cli/codex	10	n/a	n/a	n/a – telemetry unavailable	10 / 10	100.0%
mistral:codestral-latest	10	172 890	17 289	not billed	0 / 10	0.0%

Findings

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

F-01

PASS

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

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

19 of 30 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 × 11 (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

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.

Verification

RUN ID

live-20260902T001847Z

EXAM FINGERPRINT

aa41977120e6ecddb5c4fc603e8ab23cb272726a86750119969ff4dfc90e297

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

138425da59aabc18aa208c339aa41919f387e1ecb80fb725cfed8f7d6ba78713

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

SOURCE RECORD FILE

02b4f83ee15251a5808a83153c9c4590936af23c58cc94d172556e9f983d1909

sha256 of fr-20260902-6616b188.json as written beside the standalone report.

SOURCE STANDALONE REPORT

1bf300e86c51f0a627dd4da9210c742932e2b3f81cdfb430cc98abe5846c021c

sha256 of fr-20260902-6616b188.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

968faece2febdcd3a42ba22d89a7e8af8e17a40249069ac2d7675ee50c853833

sha256 of fr-20260902-6616b188.pdf.

RUN EVIDENCE BUNDLE

6616b188f015f52e157bdfb257e8d0f819e1f6cb9bcd3b024b340ac0771e0fd8

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
3e6298d8	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	180.1s
cb495fe6	cli/claude-sonnet	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	150.5s
e76e1c66	cli/claude-sonnet	2	submitted_passed	yes	Passed hidden tests	pass	1	n/a	133.3s
9e19f314	cli/claude-sonnet	3	submitted_passed	yes	Passed hidden tests	pass	1	n/a	147.9s
aaa748af	cli/claude-sonnet	4	submitted_passed	yes	Passed hidden tests	pass	1	n/a	225.9s
3e36cea5	cli/claude-sonnet	5	submitted_passed	yes	Passed hidden tests	pass	1	n/a	181.1s
6516049a	cli/claude-sonnet	6	submitted_failed	yes	Graded	fail	0	n/a	208.2s
8c122607	cli/claude-sonnet	7	submitted_passed	yes	Passed hidden tests	pass	1	n/a	110.6s
3161d1b1	cli/claude-sonnet	8	submitted_passed	yes	Passed hidden tests	pass	1	n/a	100.9s
6b722882	cli/claude-sonnet	9	submitted_passed	yes	Passed hidden tests	pass	1	n/a	159.4s
92c89eb1	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	46.0s
6d48c276	cli/codex	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	40.8s
1d32b64e	cli/codex	2	submitted_passed	yes	Passed hidden tests	pass	1	n/a	40.2s
3e53d590	cli/codex	3	submitted_passed	yes	Passed hidden tests	pass	1	n/a	40.7s
90405084	cli/codex	4	submitted_passed	yes	Passed hidden tests	pass	1	n/a	51.1s
48e359d8	cli/codex	5	submitted_passed	yes	Passed hidden tests	pass	1	n/a	36.9s
fd7561b3	cli/codex	6	submitted_passed	yes	Passed hidden tests	pass	1	n/a	39.5s
4fe31a46	cli/codex	7	submitted_passed	yes	Passed hidden tests	pass	1	n/a	43.4s
18660280	cli/codex	8	submitted_passed	yes	Passed hidden tests	pass	1	n/a	46.1s
195b8f72	cli/codex	9	submitted_passed	yes	Passed hidden tests	pass	1	n/a	39.5s
5d129402	codestral-latest	0	submitted_failed	yes	Graded	fail	1	19 250	11.5s
bd30cb85	codestral-latest	1	submitted_failed	yes	Graded	fail	1	16 813	11.2s
7cc71ab5	codestral-latest	2	submitted_failed	yes	Graded	fail	1	16 792	10.6s
d38fb2e2	codestral-latest	3	submitted_failed	yes	Graded	fail	1	16 861	14.9s
8c65b44c	codestral-latest	4	submitted_failed	yes	Graded	fail	1	16 876	11.2s
b8a625a2	codestral-latest	5	submitted_failed	yes	Graded	fail	1	16 762	10.7s
58abb85c	codestral-latest	6	submitted_failed	yes	Graded	fail	1	16 876	10.6s
94122fe4	codestral-latest	7	submitted_failed	yes	Graded	fail	1	16 759	13.5s
b0843fec	codestral-latest	8	submitted_failed	yes	Graded	fail	1	19 025	10.2s
cecbcc3b	codestral-latest	9	submitted_failed	yes	Graded	fail	1	16 876	10.2s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	11	Submitted a change; the independent hidden tests still failed.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

6616b188f015f52e157bdfb2

57e8d0f819e1f6cb9bcd3b02

4b340ac0771e0fd8

RUN ID

live-

20260902T001847Z

STARTED

2026-09-

02T00:18:47.313247

+00:00

FINISHED

2026-09-

02T00:57:34.65040

8+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	059a37157968962044d98a133f9422c6831ff452419eb4b8729b856a79cad646
rollouts/18660280-d29a-4c7f-928c-0f37b96f18e3.json	7049943d0e9a63ddf108caa49598a38a76ccde4f29d1e54cb9f3098ca7aa2e37
rollouts/195b8f72-f390-486a-a369-b5047d3bf081.json	f199086fc48f883f348848c82b55fa6cdf2dd1e415001939c9176faf4ab48b84
rollouts/1d32b64e-b2fd-4a37-a04f-1230e38ad920.json	d7ad72ef466cadbccd862d52e8b0cad772ef2c9360ccb602ab332ce976bd9e44
rollouts/3161d1b1-1be0-4bcc-bb1a-11592c0931aa.json	7590bef63572befafe69f1120525598f6f03ce59869ec21a7e9d61cf2c081ced
rollouts/3e36cea5-b3a2-49a1-8227-83d3c1f6b123.json	df73c10cd29332e01062110ea33c6df2c7e592aba53fbb2b67f58e333bd1d601
rollouts/3e53d590-9fa2-40ec-9d8d-284a6663637c.json	fbaf104694620c1643a2341422af017c17122b02d6dca2599d6ea77e27d10851
rollouts/3e6298d8-5275-4323-8970-b9d0c7709aea.json	e2483eb5171491c61186b5edda6e992b7a8fd87c794a010f76557a5ae0402813
rollouts/48e359d8-3ed5-46b1-b6a1-d6aac402eef9.json	8934223b0170fdaa59fc24e801016a92df6672d4e493d17c37f96db41875ec2d
rollouts/4fe31a46-de93-4684-bc08-eda4db9ceab0.json	3058c5a8ef15ec2dea64cc04c5d4ae747c66c55de8d0b80ac7acfbcb8524554da
rollouts/58abb85c-b63a-447e-a07b-071cdcb65b4b.json	ed4c3b6913b3df2f9098a2d6e8f384cd5b6f516e0de8999923013d55ac3ae22d
rollouts/5d129402-a899-48be-b2bd-0101eeaa04c8.json	3f98b67607190f913c03acb489561a96b6ddf351c5b17ac1687ae4fc75405016
rollouts/6516049a-dc0f-4636-9df0-9f827bb6dd1b.json	6d9ba60e40f7e85a59635d2f75675caf7d330cede1bdd8146b12c3c2a4445037
rollouts/6b722882-8ce0-4eb7-8cdd-db66e413e80f.json	2960bb9ff1bd93c878aa87ad43159daca5c38a6428c2309e78c9b79d4265e408
rollouts/6d48c276-533c-469a-92d4-738279c6027d.json	0ff66e69b4d90d13bac4ce41cae996babd1bf018df5002de15d5eb0cbc33daa7
rollouts/7cc71ab5-3ba2-4a7a-8417-7e1ac353afab.json	3d7170d35dd63c1803aaa4a090f6551d23a62fb350baabe62bf396d7b317fad8
rollouts/8c122607-d35f-45fc-ba31-68bb9a296352.json	2a699f21909127d47ac15af29455549e6e62d22635fa77162bd93d485a119c42
rollouts/8c65b44c-7a1b-4da4-9194-a66dcb79f193.json	a25b7c2c5e0e41c78185730582cc68d4aaf8142ba3812c6cef418651bfc0b6ed
rollouts/90405084-4394-484b-8d91-78909aa09787.json	f5d78417af80d56b68fb13cb84831e89b3c9d76f17e3264135cf5cc8065eff4f
rollouts/92c89eb1-4f77-4165-9b73-0b53e3bff694.json	b6432a4e97b4652b098fd7b50fc4420c5a1fd530974e40ab67aeebf088351447
rollouts/94122fe4-5f7e-453e-8ef5-41250ff398d5.json	585128f0e325a96995e9acace25db8f364a264fad9c91f227f2071e8dd7e60255
rollouts/9e19f314-772f-4a29-8bee-8405ef9196b5.json	12754796c393b4883e68e4b21b205b19042c045174b3c75795f4f0cef0bcf45a
rollouts/aaa748af-8efc-491a-98f0-ad6565cf7ab5.json	17b22a74714f16384f42d2e92609d770535a376f3cdf490a0df82024b4caed8d
rollouts/b0843fec-30f3-491b-bfb0-9b352dea69da.json	be256e7d715aa6ef3f50fcc4a0024a1e6224d536f4f76d6b1fd165cb17b9723e
rollouts/b8a625a2-6238-476f-9951-1f7b9ed2565c.json	41e5b0047d43b493b175b0518edbbd82c8e5b99cd9160d3941abed1e4d22251a
rollouts/bd30cb85-1647-4722-9b99-b469583dce5e.json	a2fe27d8be1fd255d92b707bdb439075a56ec040fff728b634774b435b9e3a652
rollouts/cb495fe6-4d25-457d-9e5a-2d830c2d1e87.json	28e67f4869f6a390518199f6dc27bbd78033140b3bcb58f49c8db39ca98f0c83
rollouts/cecbcc3b-ea8a-4b61-8b95-91bc9c831eea.json	9dec1bd6a4fb9af4c240dc72a6d1d974129f9a9c699ddf371148e2c80910cc2
rollouts/d38fb2e2-3dea-4276-abe9-5c44ef17994e.json	a0f91309e405bfb8cd93ebce214a8e5ef098ebbe7994d3e0d6972aa59bcfdfb1

ARTIFACT	SHA-256
rollouts/e76e1c66-ea48-4852-920a-dc077f09b4ab.json	a323881e38c7b2c9edf6ea34a992d305abc4ad23c77bdb3d8c934c9e3f622332
rollouts/fd7561b3-afd5-4de2-91a4-11262c5b065e.json	aec6d21e216d0d84d43edc7343309a41265f82880cd6aecb9289bbcd894d0ffc
spec	26f9634cb3113984cabf0c0f5675880645449731e9d8ed3baeb124c35e1e60f5
summary	2a5e9b681ab1c80119fb4922de9f9d7b74a1e58fb7594ceebf3f6be296cf4ba5

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.

3e6298d8 · 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
@@ -107,12 +107,28 @@
     # 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 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)
+                 item_value = item
+             else:
+                 item_value = data
+         # Apply postprocessor as push_data would when storing in parent
+         if self.postprocessor is not None:
+             result = self.postprocessor(self.path, name, item_value)
+             if result is not None:
+                 _, item_value = result
+             else:
+                 item_value = None
+
+         should_continue = self.item_callback(self.path, item_value)
+         if not should_continue:
+             raise ParsingInterrupted
+         # Reset state for the parent context without keeping a reference to
```

Showing 36 of 36 diff line(s).

cb495fe6 · cli:cli/claude-sonnet · seed 1 · 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
@@ -107,10 +107,28 @@
     # 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
+         # Simulate push_data(parent, name, item) to get the value
+         # that would be stored in the full (non-streaming) parse
+         if self.postprocessor is not None:
+             result = self.postprocessor(self.path, name, item)
+             if result is not None:
+                 _, item = result
+             else:
+                 item = None
+         if self._should_force_list(name, item):
+             item = [item]

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

Showing 34 of 34 diff line(s).

e76e1c66 · cli:cli/claude-sonnet · seed 2 · 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.
+
+         # Restore parent state (stack.pop mirrors non-streaming branch)
         if self.stack:
             self.item, self.data = self.stack.pop()
         else:
             self.item = None
             self.data = []
+
+         # Apply the same transformations the non-streaming branch uses
+         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)
+                 # Simulate what push_data(parent, name, item) would do to the value
+                 if self.postprocessor is not None:
+                     result = self.postprocessor(self.path, name, item)
+                     if result is not None:
+                         _, item = result
+             else:
+                 # text-only (or empty) path
+                 if self.postprocessor is not None:

```

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

9e19f314 · cli:cli/claude-sonnet · seed 3 · 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
@@ -107,12 +107,21 @@
     # 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 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)
+             result = item
+     else:
+         result = data
+
+     should_continue = self.item_callback(self.path, result)
+     if not should_continue:
+         raise ParsingInterrupted
+     # Reset state for the parent context without keeping a reference to

```

Showing 29 of 29 diff line(s).

aaa748af · cli:cli/claude-sonnet · seed 4 · submitted_passed

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

```

--- a/xmldict.py
+++ b/xmldict.py
@@ -107,21 +107,38 @@
     # 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 transformations as the non-streaming path
+         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)
+                 callback_item = item
+         else:

```

```

+         callback_item = data
+
+         # Apply postprocessor for the element name (as push_data would do for parent)
+         if self.postprocessor is not None:
+             result = self.postprocessor(self.path, name, callback_item)
+             if result is not None:

```

Showing 40 of 50 diff line(s); truncated.

3e36cea5 · cli:cli/claude-sonnet · seed 5 · 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 +107,38 @@
     # 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
+         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)
+                 final_value = item
+             else:
+                 final_value = data
+
+         # Replicate what push_data(parent, name, final_value) would compute
+         # (postprocessor + force_list) without actually storing in parent
+         if self.postprocessor is not None:
+             result = self.postprocessor(self.path, name, final_value)
+             if result is None:
+                 self.path.pop()

```

Showing 40 of 50 diff line(s); truncated.

6516049a · cli:cli/claude-sonnet · seed 6 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	xmltodict.py	ok	12 166
1	read_file	xmltodict.py	ok	11 706
2	invalid_tool_call	(none)	failed	107
3	run_tests	—	ok	183
4	submit	—	ok	85

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

8c122607 · cli:cli/claude-sonnet · seed 7 · submitted_passed

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

```
--- a/xmltodict.py
+++ b/xmltodict.py
@@ -107,21 +107,37 @@
     # 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.
+         # Restore parent state (same ordering as non-streaming branch)
         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
+         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)
+             result = item
+         else:
+             result = data
+
+         # Apply postprocessor for the element itself
+         if self.postprocessor is not None:
+             proc_result = self.postprocessor(self.path, name, result)
```

```

+         if proc_result is not None:
+             _, result = proc_result

```

Showing 40 of 49 diff line(s); truncated.

3161d1b1 · cli:cli/claude-sonnet · seed 8 · 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 +107,28 @@
     # 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
+         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)
+                 value = item
+             else:
+                 value = data
+             should_continue = self.item_callback(self.path, value)
+             if not should_continue:
+                 raise ParsingInterrupted
+             self.path.pop()
+             return
+         if self.stack:

```

Showing 40 of 40 diff line(s).

6b722882 · cli:cli/claude-sonnet · seed 9 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	xmldict.py	ok	12 166
1	read_file	xmldict.py	ok	11 706
2	invalid_tool_call	(none)	failed	107
3	edit_file	xmldict.py	ok	50
4	run_tests	—	ok	183
5	submit	—	ok	85

```
--- a/xmldict.py
+++ b/xmldict.py
@@ -107,21 +107,39 @@
     # 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.
+
+         # Restore parent state
         if self.stack:
             self.item, self.data = self.stack.pop()
         else:
             self.item = None
             self.data = []

+
+         # Apply the same transformations as the non-streaming branch
+         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)
+             result = item
+         else:
+             result = data
+
+         # Apply postprocessor for the element itself (mirrors push_data)
+         if self.postprocessor is not None:
+             entry = self.postprocessor(self.path, name, result)
```

Showing 40 of 51 diff line(s); truncated.

92c89eb1 · 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

STEP	TOOL	PATH	RESULT	CHARS
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 +107,28 @@
     # 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
+         # that push_data() would apply when attaching this value to its
+         # parent, but use a temporary container so streaming still drops
+         # the completed element instead of retaining it in the document.
+         container = self.push_data(None, name, item)
+         if container:
+             item = container[next(iter(container))]
+         else:
+             item = None
+
+         should_continue = self.item_callback(self.path, item)
+         if not should_continue:

```

Showing 34 of 34 diff line(s).

6d48c276 · cli:cli/codex · seed 1 · 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
@@ -107,12 +107,34 @@
     # 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:

```

```

+         # Finalize the streamed element exactly as the ordinary branch
+         # would, but use a temporary parent so the document is not
+         # accumulated in streaming mode.
+         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 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)
+                 value = item
+             else:
+                 value = data
+
+         # In a full parse, the completed value is passed through
+         # push_data when it is installed under its element name. That is
+         # where the element-level postprocessor and force_list handling
+         # occur. Build only that one-entry wrapper, then discard it.
+         wrapper = self.push_data(None, name, value)
+         if wrapper is None:
+             value = None
+         else:
+             value = next(iter(wrapper.values()))
+
+         should_continue = self.item_callback(self.path, value)
+         if not should_continue:

```

Showing 40 of 42 diff line(s); truncated.

1d32b64e · cli:cli/codex · seed 2 · submitted_passed

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

```

--- a/xmldict.py
+++ b/xmldict.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 attaching the element to its parent, but use a temporary
+         # container so streamed items are still dropped after the callback.
+         container = self.push_data(None, name, item)
+         if container is None:
+             item = None
+         else:
+             item = next(iter(container.values()))
+
+         should_continue = self.item_callback(self.path, item)
+         if not should_continue:

```

Showing 33 of 33 diff line(s).

3e53d590 · cli:cli/codex · seed 3 · 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 +107,28 @@
     # 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
+         # that push_data applies when attaching an ordinary element to its
+         # parent, but keep the temporary wrapper local so streaming does
+         # not retain the completed item.
+         wrapped_item = self.push_data(None, name, item)
+         if wrapped_item is None:
+             item = None
+         else:
+             item = next(iter(wrapped_item.values()))

```

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

Showing 34 of 34 diff line(s).

90405084 · cli:cli/codex · seed 4 · 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
@@ -107,21 +107,36 @@
     # 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.stack:
+             self.item, self.data = self.stack.pop()
+         else:
+             self.item = None
+             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
+         # use a temporary container so the streamed document is not kept.
+         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:
            raise ParsingInterrupted
-
-     # Reset state for the parent context without keeping a reference to
```

Showing 40 of 49 diff line(s); truncated.

48e359d8 · cli:cli/codex · seed 5 · 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
@@ -107,11 +107,25 @@
     # 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)
+                 emitted = self.push_data(None, name, item)
+             else:
+                 emitted = self.push_data(None, name, data)
+
+         # Build the value through the same push_data path used by ordinary
+         # parsing, so element-level postprocessing and force_list apply,
+         # but do not retain the temporary containing mapping.
+         item = (None if emitted is None
+                 else next(iter(emitted.values())))
        should_continue = self.item_callback(self.path, item)
        if not should_continue:
            raise ParsingInterrupted
```

Showing 32 of 32 diff line(s).

fd7561b3 · cli:cli/codex · seed 6 · 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 +107,28 @@
     # 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:
+                 item = self.push_data(item, self.cdata_key, data)
+         else:
+             item = data
+
+         # Apply the same element-level postprocessing and force-list logic
+         # that push_data applies when attaching a completed element to its
+         # parent, but keep it in a temporary container so streaming does
+         # not retain the completed item.
+         container = self.push_data(None, name, item)
+         if container:
+             item = next(iter(container.values()))
+         else:
+             item = None

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

```

Showing 34 of 34 diff line(s).

4fe31a46 · cli:cli/codex · seed 7 · 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 +107,24 @@
     # 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 element-level postprocessor and force_list handling in
+         # exactly the same way as the ordinary parent-attachment path, but
+         # use a temporary container so streamed items are still discarded.
+         container = self.push_data(None, name, item)
+         item = next(iter(container.values()), None)
+
+         should_continue = self.item_callback(self.path, item)
+         if not should_continue:

```

Showing 30 of 30 diff line(s).

18660280 · cli:cli/codex · seed 8 · 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
@@ -103,27 +103,6 @@

def endElement(self, full_name):
    name = self._build_name(full_name)
-    # If we just closed an item at the streaming depth, emit it and drop it
-    # 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:
-        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 = []
-        self.path.pop()
-        return
-    if self.stack:
-        data = (None if not self.data
-               else self.cdata_separator.join(self.data))
@@ -135,10 +114,26 @@
        item = self.dict_constructor()
        if item is not None:
            if data:
-                self.push_data(item, self.cdata_key, data)

```

```

+         item = self.push_data(item, self.cdata_key, data)
+     else:
+         item = data
+
+     # Build the same value that push_data would place under this

```

Showing 40 of 60 diff line(s); truncated.

195b8f72 · cli:cli/codex · seed 9 · 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 +107,23 @@
     # 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:
+                 item = self.push_data(item, self.cdata_key, data)
+                 emitted = self.push_data(None, name, item)
+             else:
+                 emitted = self.push_data(None, name, data)
+             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 29 of 29 diff line(s).

5d129402 · 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
3	run_tests	—	ok	183

STEP	TOOL	PATH	RESULT	CHARS
4	submit	—	ok	85

```

--- a/xmltodict.py
+++ b/xmltodict.py
@@ -111,6 +111,17 @@
     if item is None:
         item = (None if not self.data
                 else self.cdata_separator.join(self.data))
+
+    else:
+        if self.data:
+            if self.strip_whitespace:
+                text = self.cdata_separator.join(self.data).strip()
+            else:
+                text = self.cdata_separator.join(self.data)
+            if self._should_force_cdata(name, text):
+                if self.cdata_key not in item:
+                    item[self.cdata_key] = text
+            else:
+                item[self.cdata_key] = text

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

```

Showing 20 of 20 diff line(s).

bd30cb85 · mistral:codestral-latest · seed 1 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmltodict.py	ok	12 166
2	edit_file	xmltodict.py	ok	50
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))
+
+    # Apply postprocessor and other transformations
+    if self.postprocessor:
+        item = self.postprocessor(self.path, self.path[-1][0], item)
+    if self.strip_whitespace and isinstance(item, str):
+        item = item.strip() or None
+    if self.force_cdata and isinstance(item, str):
+        item = {self.cdata_key: item}
+    if self.force_list and isinstance(item, str):
+        item = [item]

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

```

Showing 19 of 19 diff line(s).

7cc71ab5 · mistral:codestral-latest · seed 2 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmldict.py	ok	12 166
2	edit_file	xmldict.py	ok	50
3	run_tests	—	ok	183
4	submit	—	ok	85

```
--- a/xmldict.py
+++ b/xmldict.py
@@ -111,6 +111,17 @@
     if item is None:
         item = (None if not self.data
                 else self.cdata_separator.join(self.data))
+
+    else:
+        if self.data:
+            if self.strip_whitespace:
+                text = self.cdata_separator.join(self.data).strip()
+            else:
+                text = self.cdata_separator.join(self.data)
+            if text:
+                if self._should_force_cdata(self.path[-1][0], text):
+                    item[self.cdata_key] = text
+                else:
+                    item[self.cdata_key] = text
+
     should_continue = self.item_callback(self.path, item)
     if not should_continue:
```

Showing 20 of 20 diff line(s).

d38fb2e2 · mistral:codestral-latest · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmldict.py	ok	12 166
2	edit_file	xmldict.py	ok	50
3	run_tests	—	ok	183
4	submit	—	ok	85

```
--- a/xmldict.py
+++ b/xmldict.py
@@ -111,6 +111,18 @@
     if item is None:
         item = (None if not self.data
                 else self.cdata_separator.join(self.data))
+
+    # Apply postprocessor and other transformations
+    if self.postprocessor:
+        item = self.postprocessor(self.path, name, item)
+    if self.strip_whitespace and isinstance(item, str):
+        item = item.strip() or None
+    if self.force_cdata and isinstance(item, str):
+        item = {self.cdata_key: item}
+    if self.force_list and isinstance(item, dict):
+        for key in self.force_list:
```

```

+         if key in item:
+             item[key] = [item[key]]

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

```

Showing 21 of 21 diff line(s).

8c65b44c · mistral:codestral-latest · seed 4 · submitted_failed

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

```

--- a/xmltodict.py
+++ b/xmltodict.py
@@ -111,6 +111,18 @@
     if item is None:
         item = (None if not self.data
                 else self.cdata_separator.join(self.data))

+
+    # Apply postprocessor and other transformations
+    if self.postprocessor:
+        item = self.postprocessor(self.path, self.path[-1][0], item)
+    if self.strip_whitespace and isinstance(item, str):
+        item = item.strip() or None
+    if self.force_cdata and isinstance(item, str):
+        item = {self.cdata_key: item}
+    if self.force_list and isinstance(item, dict):
+        for key in self.force_list:
+            if key in item:
+                item[key] = [item[key]]

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

```

Showing 21 of 21 diff line(s).

b8a625a2 · mistral:codestral-latest · seed 5 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmltodict.py	ok	12 166
2	edit_file	xmltodict.py	ok	50
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:
+                 if self.strip_whitespace:
+                     text = self.cdata_separator.join(self.data).strip()
+                 else:
+                     text = self.cdata_separator.join(self.data)
+                 if self._should_force_cdata(name, text):
+                     item[self.cdata_key] = text
+                 else:
+                     item[self.cdata_key] = text

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

```

Showing 19 of 19 diff line(s).

58abb85c · mistral:codestral-latest · seed 6 · submitted_failed

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

```

--- a/xmltodict.py
+++ b/xmltodict.py
@@ -111,6 +111,18 @@
         if item is None:
             item = (None if not self.data
                     else self.cdata_separator.join(self.data))

+
+         # Apply postprocessor and other transformations
+         if self.postprocessor:
+             item = self.postprocessor(self.path, name, item)
+         if self.strip_whitespace and isinstance(item, str):
+             item = item.strip() or None
+         if self.force_cdata and isinstance(item, str):
+             item = {self.cdata_key: item}
+         if self.force_list and isinstance(item, dict):
+             for key in item:
+                 if self._should_force_list(key, item[key]):
+                     item[key] = [item[key]]

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

```

Showing 21 of 21 diff line(s).

94122fe4 · mistral:codestral-latest · seed 7 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmltodict.py	ok	12 166
2	edit_file	xmltodict.py	ok	50
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).

b0843fec · mistral:codestral-latest · seed 8 · 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
3	run_tests	—	ok	183
4	submit	—	ok	85

```

--- a/xmltodict.py
+++ b/xmltodict.py
@@ -102,6 +102,36 @@
     self.data = []

    def endElement(self, full_name):
+
+        name = self._build_name(full_name)
+        # If we just closed an item at the streaming depth, emit it and drop it
+        # 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:
+            item = self.item
+            if item is None:
+                item = (None if not self.data
+                       else self.cdata_separator.join(self.data))
+
+            # Apply the same processing as in non-streaming mode
+            if self.strip_whitespace and item:
+                item = item.strip() or None
+            if item and self._should_force_cdata(name, item):
+                item = {self.cdata_key: item}
+            if self.postprocessor:
+                item = self.postprocessor(self.path, name, item)
+
+            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 = []
+         self.path.pop()
+         return
name = self._build_name(full_name)
# If we just closed an item at the streaming depth, emit it and drop it
# without attaching it back to its parent. This avoids accumulating all

```

Showing 39 of 39 diff line(s).

cecbcc3b · mistral:codestral-latest · seed 9 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	39
1	read_file	xmldict.py	ok	12 166
2	edit_file	xmldict.py	ok	50
3	run_tests	—	ok	183
4	submit	—	ok	85

```

--- a/xmldict.py
+++ b/xmldict.py
@@ -111,6 +111,18 @@
     if item is None:
         item = (None if not self.data
                 else self.cdata_separator.join(self.data))

+
+     # Apply postprocessor and other transformations
+     if self.postprocessor:
+         item = self.postprocessor(self.path, self.path[-1][0], item)
+     if self.strip_whitespace and isinstance(item, str):
+         item = item.strip() or None
+     if self.force_cdata and isinstance(item, str):
+         item = {self.cdata_key: item}
+     if self.force_list and isinstance(item, dict):
+         for key in self.force_list:
+             if key in item:
+                 item[key] = [item[key]]

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

```

Showing 21 of 21 diff line(s).

10 Exam · grounded-rag-1

vvdex.knowledge.grounded-rag-1 · record fr-20260902-2e7f4171 · record digest 157ec324a925c0b1... · harness clean · containment clean

Executive summary

MODELS TESTED 3	CERTIFIED PASSES 11 of 29 graded rollouts	SUBMITTED 29 of 29 graded rollouts that called submit	NOT GRADED — LANE ERRORS 1 of 30 attempts
---------------------------------------	--	--	--

Purpose

This report presents a controlled evaluation of 3 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

11 of 29

graded rollouts passed the independent hidden tests and were certified — 1 of 30 attempts not graded (lane errors)

11 of 29 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. 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 passed the independent hidden tests
- F-02 submitted_failed × 18 (cli/claude-sonnet, codestral-latest)
- F-03 model_api_failure × 1 (codestral-latest)

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 1500.2s.

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

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

The instrument. Exam `vvdex.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	30 (10 per model)

Where the defect came from

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

Certification evidence

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-20260902-2e7f4171
ISSUED	2026-09-02
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 23:53:46	Evaluation run started
2026-09-02 00:18:47	Evaluation run finished
2026-09-02 00:18:47	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

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

Limits

RUNTIME docker	STEP LIMIT 30	WALL-CLOCK LIMIT 7320s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 40	TOOLS OFFERED 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: 30. Graded (evaluable) rollouts: 29. 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	10	10	1 / 10	9	0	10 / 10	45.8s	Graded
Codex CLI (GPT, subscription) cli/codex	10	10	10 / 10	0	0	10 / 10	67.8s	Passed hidden tests
codestral-latest codestral-latest	10	9	0 / 9	9	1	9 / 9	18.0s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
cli/claude-sonnet	10	10	0	1	10.0%	[2%, 40%]	0–9	10.0%	30.0%	50.0%	100.0%
cli/codex	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
codestral-latest	10	9	1	0	0.0%	[0%, 30%]	0–9	0.0%	0.0%	0.0%	—

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/claude-sonnet	10.0%	[2%, 40%]	n/a	n/a	45.8s	0.0%	submitted_failed
cli/codex	100.0%	[72%, 100%]	n/a	n/a	67.8s	0.0%	none
codestral-latest	0.0%	[0%, 30%]	225 330	not billed	18.0s	0.0%	submitted_failed

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

This is not a leaderboard. Separated by this run (intervals do not overlap): cli/claude-sonnet vs cli/codex; cli/codex vs codestral-latest. Not separated (intervals overlap): cli/claude-sonnet vs codestral-latest. Below the 10-rollout floor for reading a rate as a rate: codestral-latest (9) — 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
Inspected	10 / 10	10 / 10	9 / 9
Edited	8 / 10	10 / 10	9 / 9
Ran tests	10 / 10	10 / 10	9 / 9
Submitted	10 / 10	10 / 10	9 / 9
Graded	10 / 10	10 / 10	9 / 9
Passed hidden tests	1 / 10	10 / 10	0 / 9

Deepest stage reached

MODEL	DEEPEST (ANY ATTEMPT)	TYPICAL (MOST ATTEMPTS)	HIDDEN-TEST PASSES	NOT GRADED
cli:cli/claude-sonnet	Passed hidden tests	Graded	1 / 10	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	10 / 10	0
mistral:codestral-latest	Graded	Graded	0 / 9	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	10	9 / 30	0	0 / 40	0	0	0	0
cli:cli/codex	10	14 / 30	0	0 / 40	0	0	0	0
mistral:codestral-latest	10	8 / 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	10	0	none	0 / 10	0 / 10	0.0%	0
cli:cli/codex	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:codestral-latest	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 30 rollout(s). 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	10	n/a	n/a	n/a – telemetry unavailable	1 / 10	10.0%
cli:cli/codex	10	n/a	n/a	n/a – telemetry unavailable	10 / 10	100.0%
mistral:codestral-latest	9	225 330	25 036	not billed	0 / 9	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
11 of 29 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 × 18 (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 (codestral-latest)	
Severity: LANE / INFRA	Evidence: Appendix A · Appendix C
The model endpoint failed to answer. Not a verdict about the model's ability. These rollouts are excluded from any capability reading.	
F-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
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
codestral-latest

WHAT WE WOULD LOOK AT FIRST
The lane failed, not the model; these rollouts carry no signal about capability and are labeled accordingly.

Verification

RUN ID

live-20260901T235346Z

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

157ec324a925c0b10fc8769d01f628b6143c40a00fa2dc70ea309a70e55a37a8

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

SOURCE RECORD FILE

a3e742fe73fe3667155e22dfc941bdc9e428a4bb1bd47c5c89a0454e299e7eb0

sha256 of `fr-20260902-2e7f4171.json` as written beside the standalone report.

SOURCE STANDALONE REPORT

e9ec887153f75e4b0dde10c8190ab2960cb917a2622ab5edb5c9b5bff611f66d

sha256 of `fr-20260902-2e7f4171.html` — independently verifiable with its own digest.

SOURCE STANDALONE PDF

75bde84fb5e72d3643245fbe0f18addc8adb23526f44031cd55ee0537b008f8f

sha256 of `fr-20260902-2e7f4171.pdf`.

RUN EVIDENCE BUNDLE

2e7f4171f049cced92f7a3aa7ae13b2ee211449e96938e7e6c437f30289db0ff

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
8ec5820d	cli/claude-sonnet	0	submitted_failed	yes	Graded	fail	1	n/a	25.2s
72bcf971	cli/claude-sonnet	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	51.1s
62bceaaa	cli/claude-sonnet	2	submitted_failed	yes	Graded	fail	1	n/a	52.1s
a66f85fd	cli/claude-sonnet	3	submitted_failed	yes	Graded	fail	0	n/a	36.1s
3fcba938	cli/claude-sonnet	4	submitted_failed	yes	Graded	fail	1	n/a	60.0s
70cf5fcc	cli/claude-sonnet	5	submitted_failed	yes	Graded	fail	1	n/a	56.3s
d13f1d97	cli/claude-sonnet	6	submitted_failed	yes	Graded	fail	1	n/a	53.8s
9d47ee0f	cli/claude-sonnet	7	submitted_failed	yes	Graded	fail	1	n/a	24.5s
a8cca2fc	cli/claude-sonnet	8	submitted_failed	yes	Graded	fail	0	n/a	43.8s
153192c8	cli/claude-sonnet	9	submitted_failed	yes	Graded	fail	1	n/a	55.6s
781eb674	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	76.9s
e71d8908	cli/codex	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	66.0s
2158c4f5	cli/codex	2	submitted_passed	yes	Passed hidden tests	pass	1	n/a	69.7s
8b6e987c	cli/codex	3	submitted_passed	yes	Passed hidden tests	pass	1	n/a	68.8s
fe164929	cli/codex	4	submitted_passed	yes	Passed hidden tests	pass	1	n/a	68.2s
854f7dae	cli/codex	5	submitted_passed	yes	Passed hidden tests	pass	1	n/a	65.6s
18d03c3a	cli/codex	6	submitted_passed	yes	Passed hidden tests	pass	1	n/a	65.9s
854f8510	cli/codex	7	submitted_passed	yes	Passed hidden tests	pass	1	n/a	66.5s
a111a1ee	cli/codex	8	submitted_passed	yes	Passed hidden tests	pass	1	n/a	65.7s
afdb6ff0	cli/codex	9	submitted_passed	yes	Passed hidden tests	pass	1	n/a	65.2s
ab1db90e	codestral-latest	0	submitted_failed	yes	Graded	fail	1	28 352	18.5s
6ed36c95	codestral-latest	1	submitted_failed	yes	Graded	fail	1	21 309	13.9s
496ff0a6	codestral-latest	2	submitted_failed	yes	Graded	fail	1	24 495	16.5s
e411d189	codestral-latest	3	submitted_failed	yes	Graded	fail	1	36 382	24.3s
acd6942b	codestral-latest	4	model_api_failure	no	Not graded — lane error	not graded	0	9 107	9.8s
ee5a8bdb	codestral-latest	5	submitted_failed	yes	Graded	fail	1	23 906	18.5s
3897e57e	codestral-latest	6	submitted_failed	yes	Graded	fail	1	18 536	14.1s
f23935c8	codestral-latest	7	submitted_failed	yes	Graded	fail	1	21 309	14.2s
9f5bdfb8	codestral-latest	8	submitted_failed	yes	Graded	fail	1	32 417	30.0s
f96b96b6	codestral-latest	9	submitted_failed	yes	Graded	fail	1	18 624	12.2s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	18	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

2e7f4171f049cced92f7a3aa

7ae13b2ee211449e96938e7e

6c437f30289db0ff

RUN ID

live-

20260901T235346Z

STARTED

2026-09-

01T23:53:46.828443

+00:00

FINISHED

2026-09-

02T00:18:47.040828

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	e4c4792d488475e66d9750cfca4484181c78985538f65d32605e587faad1cb4f
rollouts/153192c8-0932-4da9-bf81-111446fc32a5.json	6e633b09ebe3ee95b5cb4d387fe1356b3a02781bbd741f81e78a85a0e73728af
rollouts/18d03c3a-6629-4250-886e-c73986fdece7.json	ca658ebbd354c8f20e9cfba884d8e3098d4bbc990dfac08d14fc85ed4201f660
rollouts/2158c4f5-9a18-45cc-82e5-aa84a17c46a2.json	963bf7642556ae0d1a961a485be5acb2542e4942e1e79656bacc3463cc5d5c1e
rollouts/3897e57e-353e-402d-bb53-bc68259b1aa2.json	011530ec4616efa6f0b86d2b5a42b5aea82680b86f689f8af69217b80839ade8
rollouts/3fcb938-b19c-41d4-84fe-6ab6fec3c0a5.json	d75c04727307bfd1f50555fc0f892c36a01308a740eeef7706305cb59ce371f
rollouts/496ff0a6-9f53-4b50-8c9e-d1939301419c.json	c0ff5ca7fc1649be66606e9be02f4c8bfbec36583293f0e9aefcbd6c37c2d86c
rollouts/62bceaaa-2bf1-4067-8bef-6393b41f4d42.json	97da73eca72708cf8661d79aa0ac5351166de7f8599395f5253fdccf5627ec2b
rollouts/6ed36c95-91f9-4e3c-88da-fe17c010d584.json	5f72442b16936bf2145ef717e960d09429d418db43ad8c08de1c02dec6fec762
rollouts/70cf5fcc-7581-48bd-a1a9-53b4ce24a3a9.json	b0a5bfffdf0754ff0609c73cc67e5a019e62e0fccf0e1c4630634818ba605ceb3
rollouts/72bcf971-7d06-42c0-9d6f-7f4ab5380c76.json	c95578b9f69a937f89a18d2cf3e70049a371791d10a614a6c87b7ca3ef918796
rollouts/781eb674-aa4a-4b2f-b1bc-502f4710a4fe.json	dbeff75b715e052925a9c65e68a9ea73059a54245df5c87809b82f95aaaaa9ff
rollouts/854f7dae-d4bc-441f-be3f-fc48eb324c71.json	77ee126e557bb11b7e88c38abae6bfcd9432c27cc8e9684ae49170cc69acfa1
rollouts/854f8510-cc34-4a46-a50f-771c323e3025.json	630b147f5a7c8a9a8ddc12bb638d1f92d8d54cae19ba7ec3b49a42c479a79045
rollouts/8b6e987c-8ccd-42a0-a076-30800bf7504b.json	51a15b141bcf2a6c0d377058ed74127fb4b2d1b29389638b1ef785e2919506fa
rollouts/8ec5820d-ef88-42c9-9c7f-e5cf9923de8d.json	9ab5fbbd4ef13e7c801e7f8c99343bf26ee7bb9d3f6f38cc077863b7386742b7
rollouts/9d47ee0f-e8c3-4eb8-a5cd-ba455bfb80a3.json	21e2ecff0304fbd74fa806d478ce3e2e4140a6cf2054fc03ea99a9ab3ab96cec
rollouts/9f5bdfb8-69a0-4b45-9c77-2f87b5217ca7.json	b298cff46d8ab12505f0510a561f6ca4a2a58c6dc5ce2af3ab83586acf0728c
rollouts/a111a1ee-81a7-4118-827f-2f8cf436e16c.json	fc831c3cbaaf9dba9f5100de58cbc210c43dd42997d6dfebc088dd4ff75aff5c
rollouts/a66f85fd-827a-4d31-b407-6b0983c085ab.json	204a94e5e1d59d5faa5433a5b740546258c664c75b0e5b6490326b817bcb8aa5
rollouts/a8cca2fc-949e-4473-a79f-ac3027cab760.json	7775ac3ca099f5182d53326959308b3cd425565d5dfb7215e73f014bbe546537
rollouts/ab1db90e-5252-4819-b24b-0770692d26a9.json	d6bbea07e8570149b9f65ea3844ea0eeefc7dc5acc07253e271a2aed94b7628ed
rollouts/acd6942b-ff2c-478b-8fc5-36473631c0db.json	cdee6f5f35965fda4a7cda3d1eaab8aa1e6e22a13c0912a2afdd5811b35cc2b6
rollouts/afdb6ff0-9c18-494d-874a-43fce0ae7de7.json	dc31e5124a7738f55023ce1dea643d88d14cbfc18a1165e9003de6447be90f12
rollouts/d13f1d97-627d-4353-b467-a02faeff3156.json	b3e529077b975790be00a38c23a8f708f019c526165fb0ef68213cd2176a8731
rollouts/e411d189-390c-40ef-9318-3fbbbfb44d1.json	9fe97e4b4806ef362789ff8206bc560946bfc400b1d3046d5789534beb7c94ca
rollouts/e71d8908-6f2a-4362-9b6f-bdaf29b367f9.json	29ad3560920eadc86473f2efefdb0ee2e60c45ab030d4f68b6d95e282a8be013
rollouts/ee5a8bdb-0fe4-4909-a6f1-071b188dee55.json	d3ef142e69e1f9d296d05046099a73cb442e1df77c73463ac027717e0a749102
rollouts/f23935c8-c08a-442e-ba1d-ed39ea664b9d.json	9f3476ef0cbb5bca9e19b0e738ed19bfecb0238ef784503aa1cf0a46b9675c2f

ARTIFACT	SHA-256
rollouts/f96b96b6-e608-44d7-9f4b-8506d41d456d.json	c862c8429a774f5c5abab4442e2c2023466b8dd9eab98c117e8a1672e0209a8c
rollouts/fe164929-6cd7-4f59-acbc-ccb6ad02ddb2.json	97fc6c8b9bc91843bfa5958ffb79146d35a7c921e2f464093a7ff5736c1c6097
spec	0a225ad1c6f73d8104d73d1b5f0717e8516fc4c57f1957fc16447f8551663daa
summary	b77bc1cfc26a8c6e1848f0a23e0379695a67cf5a2db2f0543b8d6ade3a1bd1c1

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.

8ec5820d · 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	run_command	—	ok	1 780
3	write_file	answers/answers.json	ok	97
4	run_tests	—	ok	181
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (57 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.

72bcf971 · cli:cli/claude-sonnet · seed 1 · submitted_passed

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.

62bceaaa · cli:cli/claude-sonnet · seed 2 · 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.

a66f85fd · cli:cli/claude-sonnet · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	run_command	—	ok	132
3	run_command	—	ok	1 399
4	run_command	—	ok	758
5	invalid_tool_call	(none)	failed	120
6	run_tests	—	ok	181
7	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.

3fcba938 · cli:cli/claude-sonnet · seed 4 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	read_file	corpus/org-chart-2026.txt	ok	183
3	read_file	corpus/data-protection-policy-v3.txt	ok	293

STEP	TOOL	PATH	RESULT	CHARS
4	read_file	corpus/retention-addendum-2026.txt	ok	142
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.

70cf5fcc · cli:cli/claude-sonnet · seed 5 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	read_file	corpus/org-chart-2026.txt	ok	183
3	read_file	corpus/data-protection-policy-v3.txt	ok	293
4	read_file	corpus/retention-addendum-2026.txt	ok	142
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	write_file	answers/answers.json	ok	69
9	run_tests	—	ok	181
10	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.

d13f1d97 · cli:cli/claude-sonnet · seed 6 · 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

STEP	TOOL	PATH	RESULT	CHARS
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 (4 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.

9d47ee0f · cli:cli/claude-sonnet · seed 7 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	run_command	—	ok	1 780
3	write_file	answers/answers.json	ok	69
4	run_tests	—	ok	181
5	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (4 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.

a8cca2fc · cli:cli/claude-sonnet · seed 8 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	questions.json	ok	883
2	run_command	—	ok	1 399
3	run_command	—	ok	758
4	run_command	—	ok	414
5	invalid_tool_call	(none)	failed	120
6	run_tests	—	ok	181
7	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.

153192c8 · cli:cli/claude-sonnet · seed 9 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287

STEP	TOOL	PATH	RESULT	CHARS
1	read_file	questions.json	ok	883
2	read_file	corpus/data-protection-policy-v3.txt	ok	293
3	read_file	corpus/org-chart-2026.txt	ok	183
4	read_file	corpus/retention-addendum-2026.txt	ok	142
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 (4 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.

781eb674 · 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-v3.txt	ok	293
4	read_file	corpus/data-protection-policy-v1.txt	ok	239
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
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.

e71d8908 · cli:cli/codex · seed 1 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287

STEP	TOOL	PATH	RESULT	CHARS
1	read_file	ISSUE.md	ok	1 953
2	read_file	questions.json	ok	883
3	read_file	corpus/data-protection-policy-v3.txt	ok	293
4	read_file	corpus/data-protection-policy-v1.txt	ok	239
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
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 (13 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

2158c4f5 · cli:cli/codex · seed 2 · 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
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.

8b6e987c · cli:cli/codex · seed 3 · 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
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.

fe164929 · cli:cli/codex · seed 4 · 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
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.

854f7dae · cli:cli/codex · seed 5 · 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
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 (13 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

18d03c3a · cli:cli/codex · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	list_files	.	ok	287
1	read_file	ISSUE.md	ok	1 953
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
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

STEP	TOOL	PATH	RESULT	CHARS
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.

854f8510 · cli:cli/codex · seed 7 · 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
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.

a111a1ee · cli:cli/codex · seed 8 · 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
9	read_file	corpus/security-whitepaper.txt	ok	142
10	read_file	corpus/vendor-list.txt	ok	153

STEP	TOOL	PATH	RESULT	CHARS
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.

afdb6ff0 · cli:cli/codex · seed 9 · 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
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.

ab1db90e · 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	list_files	corpus	ok	249
3	read_file	corpus/data-protection-policy-v3.txt	ok	293
4	read_file	corpus/org-chart-2026.txt	ok	183
5	run_tests	—	ok	181
6	invalid_tool_call	(none)	failed	120
7	write_file	answers/answers.json	ok	69
8	run_tests	—	ok	181

STEP	TOOL	PATH	RESULT	CHARS
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.

6ed36c95 · mistral:codestral-latest · seed 1 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	answers	ok	13
3	list_files	corpus	ok	249
4	read_file	corpus/data-protection-policy-v3.txt	ok	293
5	write_file	answers/answers.json	ok	69
6	run_tests	—	ok	181
7	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.

496ff0a6 · mistral:codestral-latest · seed 2 · 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	list_files	corpus	ok	249
4	read_file	corpus/retention-addendum-2026.txt	ok	142
5	read_file	corpus/data-protection-policy-v3.txt	ok	293
6	write_file	answers/answers.json	ok	69
7	run_tests	—	ok	181
8	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.

e411d189 · mistral:codestral-latest · seed 3 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	list_files	answers	ok	13

STEP	TOOL	PATH	RESULT	CHARS
2	read_file	answers/.keep	ok	0
3	list_files	corpus	ok	249
4	read_file	corpus/data-protection-policy-v1.txt	ok	239
5	read_file	corpus/data-protection-policy-v3.txt	ok	293
6	read_file	corpus/employee-handbook.txt	ok	167
7	read_file	questions.json	ok	883
8	run_tests	–	ok	181
9	invalid_tool_call	(none)	failed	120
10	write_file	answers/answers.json	ok	69
11	run_tests	–	ok	181
12	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.

acd6942b · mistral:codestral-latest · seed 4 · model_api_failure

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	corpus	ok	249
3	invalid_tool_call	(none)	failed	120

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.

ee5a8bdb · mistral:codestral-latest · seed 5 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	answers	ok	13
3	read_file	answers/.keep	ok	0
4	list_files	corpus	ok	249
5	read_file	corpus/retention-addendum-2026.txt	ok	142
6	write_file	answers/answers.json	ok	69
7	run_tests	–	ok	181
8	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.

3897e57e · mistral:codestral-latest · seed 6 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	corpus	ok	249
3	read_file	corpus/retention-addendum-2026.txt	ok	142
4	write_file	answers/answers.json	ok	69
5	run_tests	—	ok	181
6	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.

f23935c8 · mistral:codestral-latest · seed 7 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	answers	ok	13
3	list_files	corpus	ok	249
4	read_file	corpus/data-protection-policy-v3.txt	ok	293
5	write_file	answers/answers.json	ok	69
6	run_tests	—	ok	181
7	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.

9f5bdfb8 · mistral:codestral-latest · seed 8 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	list_files	answers	ok	13
2	read_file	answers/.keep	ok	0
3	list_files	corpus	ok	249
4	read_file	corpus/data-protection-policy-v1.txt	ok	239
5	read_file	corpus/data-protection-policy-v3.txt	ok	293

STEP	TOOL	PATH	RESULT	CHARS
6	write_file	answers/answers.json	ok	69
7	run_tests	—	ok	181
8	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (15 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

f96b96b6 · mistral:codestral-latest · seed 9 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 953
1	read_file	questions.json	ok	883
2	list_files	corpus	ok	249
3	read_file	corpus/data-protection-policy-v3.txt	ok	293
4	write_file	answers/answers.json	ok	69
5	run_tests	—	ok	181
6	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.

11 Exam · memory-fact-update-1

vvdex.knowledge.memory-fact-update-1 · record fr-20260901-46eb9548 · record digest 1127e8356d42d317... · harness clean · containment clean

Executive summary

MODELS TESTED 3	CERTIFIED PASSES 22 of 30 graded rollouts	SUBMITTED 30 of 30 graded rollouts that called submit	NOT GRADED — LANE ERRORS 0 of 30 attempts
---------------------------------------	--	--	--

Purpose

This report presents a controlled evaluation of 3 model lane(s) against the certified exam vvdex.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

22 of 30

graded rollouts passed the independent hidden tests and were certified

22 of 30 graded rollouts passed. cli/claude-sonnet, cli/codex, codestral-latest completed the full loop (inspect, edit, run the visible tests, submit) and the sealed grader accepted the submission. 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, codestral-latest passed the independent hidden tests

F-02 submitted_failed × 8 (codestral-latest)

Each finding is stated in full, with evidence, in the Findings section; recommendations for acting on them follow it. Elapsed wall clock for the whole engagement: 695.7s.

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 20 tool steps and 4920s of wall clock, at 10 rollout(s) per lane. Behavioural measurements along the way (tool discipline, grounding, overclaim) are in scope because they are read from the recorded trace, not judged by another model.

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

The instrument. Exam `vvdex.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	30 (10 per model)

Where the defect came from

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

Certification evidence

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-20260901-46eb9548
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 23:42:10	Evaluation run started
2026-09-01 23:53:46	Evaluation run finished
2026-09-01 23:53:46	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

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

Limits

RUNTIME docker	STEP LIMIT 20	WALL-CLOCK LIMIT 4920s	MAX OUTPUT TOKENS 2048
TEMPERATURE 0.2	RETRY POLICY none	COMMAND BUDGET 24	TOOLS OFFERED 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: 30. Graded (evaluable) rollouts: 30. Not graded — lane errors: 0. A lane error is a rollout the API or the runtime ended before the hidden grader ran: it is listed, excluded from every rate and pass@k, and never silently dropped. Passes are certified passes: hidden grader exit 0 with integrity verified.

MODEL	ATTEMPTS	GRADED	PASSED	MODEL FAILS	LANE ERRORS	SUBMITTED	MEAN TIME	TYPICAL DEEPEST STAGE
Claude Code CLI (Sonnet, subscription) cli/claude-sonnet	10	10	10 / 10	0	0	10 / 10	17.0s	Passed hidden tests
Codex CLI (GPT, subscription) cli/codex	10	10	10 / 10	0	0	10 / 10	26.3s	Passed hidden tests
codestral-latest codestral-latest	10	10	2 / 10	8	0	10 / 10	7.4s	Graded

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

Success rate, confidence and pass@k

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

MODEL	ATTEMPTS	N	NOT GRADED	PASSES	RATE	95% INTERVAL	SEEDS	PASS@1	PASS@3	PASS@5	PASS@10
cli/claude-sonnet	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
cli/codex	10	10	0	10	100.0%	[72%, 100%]	0–9	100.0%	100.0%	100.0%	100.0%
codestral-latest	10	10	0	2	20.0%	[6%, 51%]	0–9	20.0%	53.3%	77.8%	100.0%

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

Model comparison

MODEL	PASS@1	95% INTERVAL	TOKENS	COST	MEAN ROLLOUT	OVERCLAIM RATE	TOP FAILURE CLASS
cli/claude-sonnet	100.0%	[72%, 100%]	n/a	n/a	17.0s	0.0%	none
cli/codex	100.0%	[72%, 100%]	n/a	n/a	26.3s	0.0%	none
codestral-latest	20.0%	[6%, 51%]	60 737	not billed	7.4s	0.0%	submitted_failed

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

This is not a leaderboard. Separated by this run (intervals do not overlap): cli/claude-sonnet vs codestral-latest; cli/codex vs codestral-latest. Not separated (intervals overlap): cli/claude-sonnet vs cli/codex. 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
Inspected	10 / 10	10 / 10	10 / 10
Edited	10 / 10	10 / 10	10 / 10
Ran tests	10 / 10	10 / 10	10 / 10
Submitted	10 / 10	10 / 10	10 / 10
Graded	10 / 10	10 / 10	10 / 10
Passed hidden tests	10 / 10	10 / 10	2 / 10

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	10 / 10	0
cli:cli/codex	Passed hidden tests	Passed hidden tests	10 / 10	0
mistral:codestral-latest	Passed hidden tests	Graded	2 / 10	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	10	5 / 20	0	0 / 24	0	0	0	0
cli:cli/codex	10	6 / 20	0	0 / 24	0	0	0	0
mistral:codestral-latest	10	4 / 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	10	0	none	0 / 10	0 / 10	0.0%	0
cli:cli/codex	10	0	none	0 / 10	0 / 10	0.0%	0
mistral:codestral-latest	10	0	none	0 / 10	0 / 10	0.0%	0

No grounding failures and no false pass claims were detected across these 30 rollout(s). 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	10	n/a	n/a	n/a — telemetry unavailable	10 / 10	100.0%
cli:cli/codex	10	n/a	n/a	n/a — telemetry unavailable	10 / 10	100.0%
mistral:codestral-latest	10	60 737	6 073	not billed	2 / 10	20.0%

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, codestral-latest passed the independent hidden tests

Severity: PASS

Evidence: Model outcomes · Stage funnel · Appendix A

22 of 30 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 × 8 (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

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.

Verification

RUN ID

live-20260901T234210Z

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

1127e8356d42d31767a3ddc4d6c6cd0a515f5bbf7239ef0a82129e00fc2e6900

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

SOURCE RECORD FILE

d3f91221afd12e368b53c09a12eba41633c305d41855842ca6eef918de09e857

sha256 of fr-20260901-46eb9548.json as written beside the standalone report.

SOURCE STANDALONE REPORT

956735c0334135e15df15d8b9bfab36522bddb6257950d928234f86d855f6035

sha256 of fr-20260901-46eb9548.html — independently verifiable with its own digest.

SOURCE STANDALONE PDF

68483ecfa36aded13f3f58c9e0dc2307f12f09deb0fe936fc3e1509fed7fba87

sha256 of fr-20260901-46eb9548.pdf.

RUN EVIDENCE BUNDLE

46eb954840000beb6e8466f409bb6104c91cd91a0d34efa18a41871246efd142

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
c3e59f84	cli/claude-sonnet	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	20.5s
3f5d33eb	cli/claude-sonnet	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	16.5s
91cd84ed	cli/claude-sonnet	2	submitted_passed	yes	Passed hidden tests	pass	1	n/a	15.7s
d3ccb200	cli/claude-sonnet	3	submitted_passed	yes	Passed hidden tests	pass	1	n/a	14.7s
a41f48cc	cli/claude-sonnet	4	submitted_passed	yes	Passed hidden tests	pass	1	n/a	18.9s
be0bb482	cli/claude-sonnet	5	submitted_passed	yes	Passed hidden tests	pass	1	n/a	15.6s
295bfb44	cli/claude-sonnet	6	submitted_passed	yes	Passed hidden tests	pass	1	n/a	15.8s
dd33c07a	cli/claude-sonnet	7	submitted_passed	yes	Passed hidden tests	pass	1	n/a	18.3s
c138c3a9	cli/claude-sonnet	8	submitted_passed	yes	Passed hidden tests	pass	1	n/a	15.1s
9d3daa5f	cli/claude-sonnet	9	submitted_passed	yes	Passed hidden tests	pass	1	n/a	18.8s
37ff761f	cli/codex	0	submitted_passed	yes	Passed hidden tests	pass	1	n/a	26.7s
7452d373	cli/codex	1	submitted_passed	yes	Passed hidden tests	pass	1	n/a	26.2s
9d7733ce	cli/codex	2	submitted_passed	yes	Passed hidden tests	pass	1	n/a	29.9s
83cc4801	cli/codex	3	submitted_passed	yes	Passed hidden tests	pass	1	n/a	26.9s
607bc653	cli/codex	4	submitted_passed	yes	Passed hidden tests	pass	1	n/a	26.5s
f495da98	cli/codex	5	submitted_passed	yes	Passed hidden tests	pass	1	n/a	25.9s
e880dfa5	cli/codex	6	submitted_passed	yes	Passed hidden tests	pass	1	n/a	25.0s
730a0c3f	cli/codex	7	submitted_passed	yes	Passed hidden tests	pass	1	n/a	24.3s
59a40ede	cli/codex	8	submitted_passed	yes	Passed hidden tests	pass	1	n/a	25.8s
1b564187	cli/codex	9	submitted_passed	yes	Passed hidden tests	pass	1	n/a	26.3s
d38a9234	codestral-latest	0	submitted_failed	yes	Graded	fail	1	5 764	7.2s
0f23b24f	codestral-latest	1	submitted_failed	yes	Graded	fail	1	5 764	7.2s
17fd7879	codestral-latest	2	submitted_failed	yes	Graded	fail	1	5 764	6.6s
cee5d0af	codestral-latest	3	submitted_passed	yes	Passed hidden tests	pass	1	7 279	8.1s
2e244621	codestral-latest	4	submitted_failed	yes	Graded	fail	1	5 800	7.1s
d85946ef	codestral-latest	5	submitted_failed	yes	Graded	fail	1	5 764	6.6s
1ba66ffd	codestral-latest	6	submitted_failed	yes	Graded	fail	1	5 749	6.4s
217e2513	codestral-latest	7	submitted_failed	yes	Graded	fail	1	5 735	6.9s
4f72002b	codestral-latest	8	submitted_passed	yes	Passed hidden tests	pass	1	7 354	10.0s
036f7a4e	codestral-latest	9	submitted_failed	yes	Graded	fail	1	5 764	7.9s

Failure taxonomy

CLASS	COUNT	WHAT IT MEANS
submitted_failed	8	Submitted a change; the independent hidden tests still failed.

Appendix B — evidence and artifact digests

RUN BUNDLE DIGEST

46eb954840000beb6e8466f4

09bb6104c91cd91a0d34efa1

8a41871246efd142

RUN ID

live-

20260901T234210Z

STARTED

2026-09-

01T23:42:10.844945

+00:00

FINISHED

2026-09-

01T23:53:46.554546

+00:00

Artifact digests

ARTIFACT	SHA-256
publicSummary	f00eecfd8c1b1f59dba8b9a50ea2f63d453b4e0510a0c539a07f29b817333639
rollouts/036f7a4e-95a8-44f4-9416-4a630956bde5.json	d078da23ce6dd43330d34fbca2f32ae906d034258b3aa6ad54082119bffcbbfd4
rollouts/0f23b24f-8799-4da2-8eed-9409735afdfc.json	1fcf3e129ae41a02f38af0d9a98f4363fa1440c365311bef7fac89e35c98d70f
rollouts/17fd7879-a548-4888-b365-393483b721f2.json	e96c42e097096a493d9b8a62c847c7ae08199ace0b64a9c0185742753a99e82b
rollouts/1b564187-de5c-4b9e-bba0-a3497ba2abc2.json	820522a65667acba036bb85c75e50022864897aeadf268460be1a69015d233a4
rollouts/1ba66ffd-229b-40e8-b3b0-2dae920c85b1.json	54313858ba041c81401c4e7730dca90280991e21d7014a191ba8a9dedba0ef2b
rollouts/217e2513-92b3-4f09-9148-da9988b90f7f.json	2405e3e91b701f9be6136075080b29913d9befbd38e5e7293664929941eff90c
rollouts/295bfb44-12e8-48f8-8bdc-f5df9bcecc9b.json	470e991bd1b3c34e333f75659a78b7766c90c4f98c1d01cdd5e61809ff2cb042
rollouts/2e244621-3ca6-446e-9945-ed98537cc81b.json	7b786e0eaa7fa5202436cd67418fc96bec6068b29d0b9a30f53b253a1920894c
rollouts/37ff761f-a2b3-41fa-9b42-26ed00fb4d26.json	0ab83298fae4aca11a763554a35211137c3e0d1fbe46d59466c5a04bfcacf5c43
rollouts/3f5d33eb-55bb-4320-b67c-2c20616d2509.json	0c25ae80d7162cdc131aa3168c52798a84cd18c33ec974100028dcd9cd6d09a4
rollouts/4f72002b-67f3-42b3-928a-79d7022d5180.json	c08d218f7332e24e3008c9c55814e5e29923041a02082d90f7c80defe1e741b3
rollouts/59a40ede-e515-48e5-9aa7-cce5bdd309b1.json	b33d19ff66eac735813648cfd97eeda7dd2de54eb8b82ad3b415cec658cc8f0d
rollouts/607bc653-711f-45d6-a26b-7bdfd01b0142.json	6bba74788ab8b38728a1fb1a72f07888544667fc5052389a6db6add67d782f1f
rollouts/730a0c3f-2c1b-43a0-a6c7-f005a0c7a0d6.json	c31b768016f3e9260afc4c7c8478eecdd4406a35377cf4340a49ee3514eaa5eb8
rollouts/7452d373-a07f-4c66-951b-b8c82c82badd.json	16725edfff8953f43b792cf66a786bf628a0a0e1e062413fa2f6d8327357e420
rollouts/83cc4801-1d9e-432d-81b1-31caa05f6033.json	237bfa5642324e6eac921a7aa6740045e8159a3604c9349dfec9c3551f07ec0c
rollouts/91cd84ed-bdd4-4eb6-9de6-56238bcb1d17.json	05848a847f9a10cb04d4b458675838967c5f43d7ff3dd3f19b23358ff4e0b3ce
rollouts/9d3daa5f-a2fe-4469-87dc-539a6d2d3d45.json	0182e362994ab161f4cf7ce4ed1a5206c64ad1c21a847412635538c8864d8e1b
rollouts/9d7733ce-0c7c-4f9f-8491-2d58b27af9d7.json	872aca186137b7f25f1454c455310ab41a76481ea656e1c49dfccb7b9f21e97d
rollouts/a41f48cc-c287-42e1-81b0-152f4f2017e6.json	a16e127ec3f8e19a5a816a9ae53fd4f343336e65efbb0de94fe28991caacd5b5
rollouts/be0bb482-0aa0-4452-aebc-78d3267da5c6.json	e9969415dbc882c8c8d1a53b1cc7b23adf2525700c97cbdb046f6f291691f10e
rollouts/c138c3a9-d1fc-4396-8437-dfa9356e8f57.json	ffe83adba7a97ad3ef5b2f83a52968f7b382ad2e88f252464e060e78c319edfe
rollouts/c3e59f84-2de4-4f17-9165-32d58a341c89.json	6a323bc275c391f285dd1bb3c2c67cd504190eef3f19636da5339f2caccc16b2
rollouts/cee5d0af-4f81-49fa-9da2-3ec9f3f4c2b1.json	0941fecc01b8577f8aaa411cd47566e51749690ac3280ec036282c6d1753fd65
rollouts/d38a9234-5c00-4282-b6d1-8597b2cbd8b5.json	d89c7fa3b927cb83fac42b3cadf115e0232420df1dd9675804079c2251e5b883
rollouts/d3ccb200-349b-4f7d-83fb-8cdb6453a294.json	d8ffbb6990497de81d41d9c34e29bfcfb6665dba0dfe2ee00c4df4411b1db866f
rollouts/d85946ef-5125-4b10-806a-9d17237fed3f.json	2418001521dae6707aa7154745363f2e55f589a1fbcb04ff66eeda00c924a16a
rollouts/dd33c07a-e79c-4cfa-9367-eb1c99991173.json	62da7725d6f6ee0b5c479b7336dfa4d1aaee713977a6df765ac506f5e0585479

ARTIFACT	SHA-256
rollouts/e880dfa5-8f79-4a13-817f-5be17c6dfba7.json	b33a34cf8dc3133a80c45e33a78b1d8e386c03fe8d9f0f253143db986ae57ddc
rollouts/f495da98-b293-4bd7-8722-bfb4966bc3ce.json	f2a4ffc76691fa6e27d96a489f2410c0c649f91fbd0d41307ae46cbf5abf1d1f
spec	cdaaa1f547234167c055e49c22f93ad3965c87c32220a4bf8c4c7cf94ea5ce50
summary	d5fb8760400394f5e669eff1476bdfccb37e0140a48e65d4f6954001d4cafd99

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.

c3e59f84 · 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.

3f5d33eb · cli:cli/claude-sonnet · seed 1 · 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.

91cd84ed · cli:cli/claude-sonnet · seed 2 · 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.

d3ccb200 · cli:cli/claude-sonnet · seed 3 · 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	96
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.

a41f48cc · cli:cli/claude-sonnet · seed 4 · 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.

be0bb482 · cli:cli/claude-sonnet · seed 5 · 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.

295bfb44 · cli:cli/claude-sonnet · seed 6 · 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.

dd33c07a · cli:cli/claude-sonnet · seed 7 · 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.

c138c3a9 · cli:cli/claude-sonnet · seed 8 · 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.

9d3daa5f · cli:cli/claude-sonnet · seed 9 · 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

STEP	TOOL	PATH	RESULT	CHARS
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.

37ff761f · 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.

7452d373 · cli:cli/codex · seed 1 · 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.

9d7733ce · cli:cli/codex · seed 2 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	list_files	.	ok	57
2	read_file	facts/current.json	ok	65
3	read_file	memory/notes.md	ok	135
4	write_file	answers/answer.json	ok	68
5	run_tests	—	ok	181

STEP	TOOL	PATH	RESULT	CHARS
6	submit	—	ok	85

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

83cc4801 · cli:cli/codex · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	list_files	.	ok	57
2	read_file	facts/current.json	ok	65
3	read_file	memory/notes.md	ok	135
4	write_file	answers/answer.json	ok	68
5	run_tests	—	ok	181
6	submit	—	ok	85

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

607bc653 · cli:cli/codex · seed 4 · 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.

f495da98 · cli:cli/codex · seed 5 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	read_file	memory/notes.md	ok	135
2	read_file	facts/current.json	ok	65
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.

e880dfa5 · cli:cli/codex · seed 6 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	read_file	memory/notes.md	ok	135
2	read_file	facts/current.json	ok	65
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.

730a0c3f · cli:cli/codex · seed 7 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	read_file	memory/notes.md	ok	135
2	read_file	facts/current.json	ok	65
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.

59a40ede · cli:cli/codex · seed 8 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	ISSUE.md	ok	1 220
1	read_file	memory/notes.md	ok	135
2	read_file	facts/current.json	ok	65
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.

1b564187 · cli:cli/codex · seed 9 · 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.

d38a9234 · 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 (4 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.

0f23b24f · mistral:codestral-latest · seed 1 · 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 (4 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.

17fd7879 · mistral:codestral-latest · seed 2 · 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 (4 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.

cee5d0af · mistral:codestral-latest · seed 3 · submitted_passed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	facts/current.json	ok	65
1	read_file	memory/notes.md	ok	135
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 (4 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.

2e244621 · mistral:codestral-latest · seed 4 · 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.

d85946ef · mistral:codestral-latest · seed 5 · 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 (4 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.

1ba66ffd · mistral:codestral-latest · seed 6 · submitted_failed

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

STEP	TOOL	PATH	RESULT	CHARS
2	run_tests	—	ok	181
3	submit	—	ok	85

Submission diff withheld by the public-disclosure policy (4 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.

217e2513 · mistral:codestral-latest · seed 7 · submitted_failed

STEP	TOOL	PATH	RESULT	CHARS
0	read_file	facts/current.json	ok	65
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.

4f72002b · mistral:codestral-latest · seed 8 · submitted_passed

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

036f7a4e · mistral:codestral-latest · seed 9 · 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 (4 line(s) changed). Submission content withheld — proprietary evaluation material. On this exam the submitted change is the graded answer, so publishing an excerpt of it would publish the answer key.

12 Evidence

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

- `public.swe.martinblech-xmltodict-issue-257` → `fr-20260902-6616b188` · `recordDigest` `138425da59aabc18aa208c339aa41919f387e1ecb80fb725cfed8f7d6ba78713`
- `vvdex.knowledge.grounded-rag-1` → `fr-20260902-2e7f4171` · `recordDigest` `157ec324a925c0b10fc8769d01f628b6143c40a00fa2dc70ea309a70e55a37a8`
- `vvdex.knowledge.memory-fact-update-1` → `fr-20260901-46eb9548` · `recordDigest` `1127e8356d42d31767a3ddc4d6c6cd0a515f5bbf7239ef0a82129e00fc2e6900`

13 Attestation

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

© VVDex. All rights reserved. Public materials are provided for viewing, evaluation and verification. Reuse, redistribution, derivative benchmark creation, training use, republication or commercial use requires written permission unless a specific file states otherwise. Full terms: vvdexops.com/terms/. Third-party open source reproduced in an exam keeps its own licence.