

Five exams, five families — ten rollouts per lane

5 sealed evaluations · 148 graded rollouts · 93
certified passes

CURRENT

CAMPAIGN REPORT — EXECUTIVE EDITION

2026-09-02

Evaluation scope:

cli/claude-sonnet

cli/codex

codestral-latest

Prepared by:

VVDex Forge — regenerated from verified sealed records only

fc-d89e429d2781 · 5 record(s) · every value recomputable

Contents

01.	Executive summary	3
02.	Per-exam results	5
03.	Environment families	6
04.	Campaign aggregate across these five exams	7
05.	Harness status	8
06.	Cells excluded from capability	9
07.	What the outcomes mean	10
08.	What this report withholds, and why	11
09.	Evidence	13
10.	Attestation	14

01 Executive summary

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

EXAMS 5	SEALED RECORDS 5	MODEL LANES 3
ATTEMPTS 150	GRADED ROLLOUTS 148	CERTIFIED PASSES 93
LANE ERRORS (NOT GRADED) 2	WITHHELD / INVALID 0	

Purpose

This report presents a controlled evaluation of 3 model lane(s) against 5 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. The executive edition omits record-level chapters; the full report renders each sealed evaluation record as a chapter derived from the same record as its standalone report. The full report authenticates the chapter bytes, while the standalone report carries its own digest.

Outcome

93

certified pass(es) across 148 graded rollouts (150 attempts, 2 lane error(s) not graded) in 15 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.

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).
- fault-recovery-1 · cli/claude-sonnet · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- fault-recovery-1 · cli/codex · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- fault-recovery-1 · codestral-latest · **ATTENTION** — 0 certified passes / 10 graded rollouts; 10 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- order-desk-1 · cli/claude-sonnet · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- order-desk-1 · cli/codex · **PASS** — 10 certified passes / 10 graded rollouts; 0 model failures; 0 excluded (0 lane errors, 0 withheld, 0 invalid).
- order-desk-1 · codestral-latest · **ATTENTION** — 1 certified passes / 9 graded rollouts; 8 model failures; 1 excluded (1 lane error, 0 withheld, 0 invalid).

02 Per-exam results

MODEL LANE	SWE PUBLIC.S WE.MARTI NBLECH-X MLTODICT -ISSUE-2 57	RAG VVDEX.KNOWLED GE.GROUNDED-R AG-1	MEMORY VVDEX.KNOW LEDGE.MEMO RY-FACT-UP DATE-1	HARNESS VVDEX.HAR NESS.FAUL T-RECOVER Y-1	BROWSER VVDEX.BROWSE R.ORDER-DESK- 1
cli/codex	10/10	10/10	10/10	10/10	10/10
cli/claude-sonnet	9/10	1/10	10/10	10/10	10/10
codestral-latest	0/10	0/9 + 1 lane error	2/10	0/10	1/9 + 1 lane error

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

03 Environment families

This campaign covers five exam(s) drawn from five 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)	<code>public.swe.martinblech-xmltodict-issue-257</code>	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)	<code>vvdex.knowledge.grounded-rag-1</code>	Grounded question answering over a small document collection.
Cross-session memory update	<code>vvdex.knowledge.memory-fact-update-1</code>	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.
Tool and harness fault recovery	<code>vvdex.harness.fault-recovery-1</code>	A code-fix task carried out through a deliberately unreliable tool loop.
Browser and computer use	<code>vvdex.browser.order-desk-1</code>	A task inside a real web application, driven through a real headless browser.

04 Campaign aggregate across these five exams

MODEL LANE	CERTIFIED PASSES	GRADED ROLLOUTS	LANE ERRORS	WITHHELD	INVALID	CERTIFIED RATE (95% CI)
cli/codex	50	50	0	0	0	100% [93-100]
cli/claude-sonnet	40	50	0	0	0	80% [67-89]
codestral-latest	3	48	2	0	0	6% [2-17]

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
order-desk-1	codestral-latest	2	lane error	The read operation timed out
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 <code>https://github.com/martinblech/xmltodict</code> 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>order-desk-1</code>	<code>state</code>	Applied consistently to 1 sealed records. state disclosure (ownership <code>vvdex_proprietary</code> — engine registry; family <code>browser</code>): the exam is VVDex evaluation IP and the graded output is the application's own state and audit trail, so no submission content, diff or state is published.
<code>fault-recovery-1</code>	<code>answer_key</code>	Applied consistently to 1 sealed records. answer-key disclosure (ownership <code>vvdex_proprietary</code> — engine registry; family <code>harness_eval</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>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 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.browser.order-desk-1` → `fr-20260902-6e618c05` · `recordDigest 59f94854d78f7ce174866ff114505b2b41ca2333f287f40b604fbb38cc267da5`
- `vvdex.harness.fault-recovery-1` → `fr-20260902-a6f12b64` · `recordDigest d5a8264aa0c20dd67c14879856a744e2ff1ab79da35ea8bcea92b4ba1ea14d5f`
- `vvdex.knowledge.grounded-rag-1` → `fr-20260902-2e7f4171` · `recordDigest 157ec324a925c0b10fc8769d01f628b6143c40a00fa2dc70ea309a70e55a37a8`
- `vvdex.knowledge.memory-fact-update-1` → `fr-20260901-46eb9548` · `recordDigest 1127e8356d42d31767a3ddc4d6c6cd0a515f5bbf7239ef0a82129e00fc2e6900`

10 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.