PURETENSOR
INTELLIGENCE DISTILLED

Quotes Are Cheap, Confidence Is Compressed

Four vendors review our code, and one reviews what the last wave never reached

Paper: PT-R-2026-071 v1.0

Author: PureTensor Ltd

Date: 8 November 2026

Status: Published

Abstract

In late October and early November 2026 we ran second-vendor adversarial code review on our own production code, using models from OpenAI, Google, xAI and Anthropic. Every finding had to carry a priority, a stated confidence, a verbatim quote of the flagged code, a scenario and a runnable reproducer. A scorer checked the quote, and an independent step checked each fix by running its new test against the unfixed and fixed code.

The quotes were a solved problem. Of 1,163 findings from the waves that recorded the check, 1,162 quoted code that exists in the source. That shows the reviewer read the file, not that the finding is real. For the two vendors whose confidence was nearly constant (OpenAI and Google), stated confidence could not separate findings, because it barely varied; we did not test it against verdicts for them, and we have no confidence data for the xAI wave. Of 733 findings from one OpenAI model across four waves, 732 were stated at 0.95 or above, including all 129 of its top-severity findings. In a separate, earlier wave, the one finding we refuted outright from that vendor was a top-severity finding at confidence 1.00 whose reproducer exercised a code path production does not use. Of the three vendors with recorded confidence, only the Anthropic models spread theirs out, and even there the link to verification was weak.

On 8 November an Anthropic model reviewed part of the code an earlier OpenAI wave had run out of allowance before reaching: 95 units, 193 findings, only 5 of them top severity. Because the reviewers ranked cautiously, we verified every second-tier finding as well, and 61 of 77 verdicts were confirmed (plus one confirmed by a second verifier). Sixty-three fixes followed; an independent check passed 51 first time. Nine of the 12 failures were a stale version number after the main branch moved; three were real defects in the fix. In a separate xAI wave, 21 percent of the commits that landed carried a follow-up in the subject, an upper bound on the defects earlier fixes created.

We cannot rank the vendors. No two reviewed the same code, and every result is a single run.

1. Why We Ran the Same Practice Across Four Vendors

We use model review as a standing practice on our own code. Each wave used a different vendor, partly because each vendor's allowance resets on its own schedule and partly to see

whether a second model catches what the first misses. The OpenAI and Google waves ended when the allowance for the window ran out, not when the work did. The OpenAI wave of 4 November, for example, reviewed 57 of 197 queued units. Only the Anthropic wave of 8 November covered units an earlier wave had not reached, and only part of them.

That arrangement lets us ask four things: how often a quoted finding quotes real code, and how often the finding itself is real; whether stated confidence says which findings to trust; whether reproducers are right; and what happens when other agents fix the findings. It cannot tell us which vendor is best. The units, briefs and code differ in every wave, and the only paired comparison is two models from one vendor on the same repository (Section 3).

2. How a Wave Works

Each wave cuts the target code into review units, each a bounded set of files. Units run in parallel in a read-only sandbox. Every reviewer receives the same adversarial brief and a JSON schema: a priority from P0 to P3, a stated confidence from 0 to 1, a file and line range, a verbatim quote, a mechanism, a scenario and a reproducer.

The scorer opens the cited file and checks that the quoted text is in it. That checks the quote, not the claim.

Serious findings go to fix agents, one per cluster of files, each in its own worktree, each writing a failing test first. An independent step resets the changed non-test sources to the base, confirms the new test fails there for the finding's reason, and confirms it passes on the branch. We call this red/green. The orchestrator, an Anthropic model in a coding harness, reruns it rather than accepting an agent's report.

Wave	Reviewer	Units reviewed	Time
26 Oct, one monitoring-configuration repository	GPT-6 Astra and GPT-6.1 Sol, extra-high reasoning, 8 units each	16	25 min, including one 13-minute retry
28 Oct, scripts repository plus one small one	GPT-6 Astra	16	10 min
30 Oct, one service repository	GPT-6 Astra	12	10 min
3 Nov, ten repositories	GPT-6 Astra	41 of 76 queued	final restart 6 min
4 Nov, four repositories	GPT-6 Astra	57 of 197 queued	8 min
31 Oct to 1 Nov, three windows	Gemini 4 Argon and Claude Opus 5.5, high reasoning, through an agent CLI in a filesystem fence	111	window 1: 36 min
31 Oct	Grok 4.7, extra-high reasoning	49 code, 48 verify, 30 fix-up, 7 research	about 3 hours of dispatch
8 Nov, four repositories (findings from two)	Claude Opus, 95 review agents	95 (662 files)	126 agents in total

Table 1. The waves. Capacity is shown only as units reviewed before the allowance ran out, and minutes.

The 3 November wave's earlier phase is not timed in the retained log, and for two of the four 8 November repositories reviewers read only the changes since the previous review.

3. Quotes Are Cheap

Reviewer	Findings	P1	P2	P3	Quote found in source
GPT-6 Astra, monitoring repo (26 Oct)	34	28	6	0	34 of 34
GPT-6.1 Sol, same repo	28	21	7	0	28 of 28
GPT-6 Astra, four later waves	733	129	575	29	733 of 733
Gemini 4 Argon, window 1	116	31	71	14	116 of 116
Claude Opus 5.5, window 1	25	1	12	12	25 of 25
Gemini 4 Argon, windows 2 and 3	216	58	132	26	215 of 216
Claude Opus 5.5, window 3	11	1	5	5	11 of 11
Claude Opus, 8 Nov, 95 units	193	5	69	119	check not recorded as a flag

Table 2. Findings by reviewer and stated priority, with the quote check.

Across every wave that recorded the flag, 1,162 of 1,163 findings (62 monitoring, 733 later OpenAI, 368 Google and Anthropic windows) quoted code that is in the file. The single miss, in window 2, quoted text that does not appear in the cited file. The 8 November findings carry no such flag and are left out of that count; the P2 verifiers reported the quote matched in all 77 verdicts.

A verbatim quote is easy to produce and easy to check, and it is no evidence that the claim built on it is true. At least one of the 77 P2 verdicts on 8 November noted that the quoted text was byte-identical to the source but the stated line numbers were wrong: it cited lines 463 to 467 in a file of 291 lines.

The only place two reviewers saw the same code was the paired comparison. On the monitoring repository, Astra and Sol reviewed the same eight units, and their 62 findings reduced to 29 distinct issues. Twenty-two were found by both models, 5 by Astra alone and 2 by Sol alone.

4. Confidence Is Compressed

Reviewer	Findings	Min	Mean	Max	At or above 0.90	At or above 0.95
Astra, monitoring repo	34	0.98	0.994	1.00	34	34
Sol, monitoring repo	28	0.97	0.992	1.00	28	28
Astra, four later waves	733	0.90	0.992	1.00	733	732
Argon, window 1	116	0.90	0.951	0.99	116	78
Argon, windows 2 and 3	216	0.88	0.954	0.99	215	150
Opus 5.5, window 1	25	0.50	0.726	0.90	2	0
Opus 5.5, window 3	11	0.35	0.554	0.85	0	0
Opus, 8 Nov	193	0.12	0.493	0.90	3	0

Table 3. Stated confidence by reviewer. Opus 5.5 is the model run through the agent CLI; Opus on 8 November is a later review in a coding harness whose exact build was not recorded in the artefacts we kept.

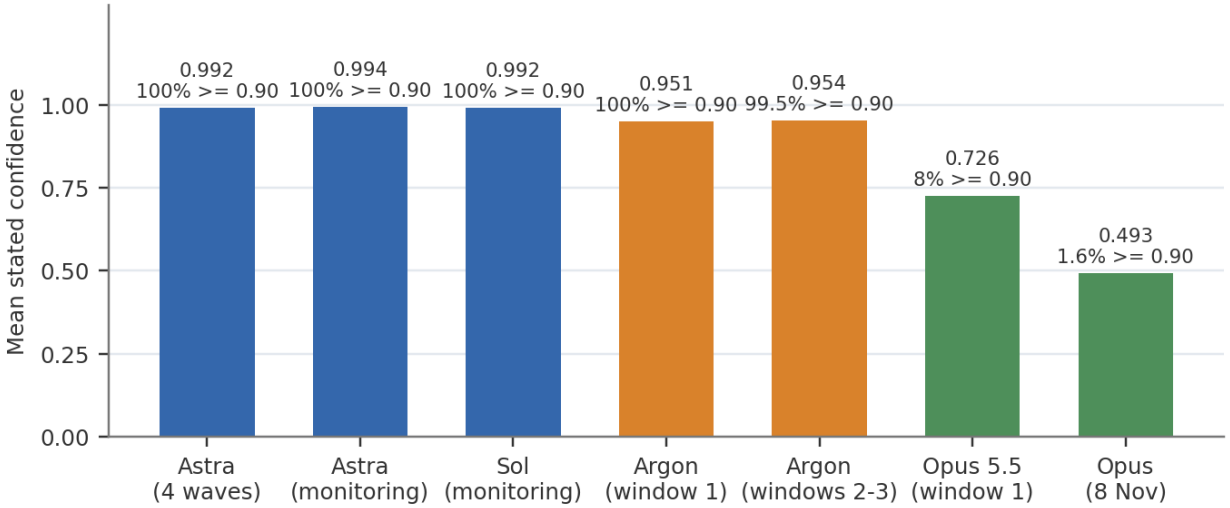


Figure 1. Mean stated confidence (bars) and the share of findings stated at 0.90 or above (labels), by reviewer. Opus 5.5 window 3 (11 findings, mean 0.554) is omitted for space. Means and counts are in Table 3; shares are the Table 3 counts at or above 0.90 divided by findings.

The two OpenAI models were nearly constant and Argon was compressed into 0.88 to 0.99, as if confidence were a house style. The scale was defined by our brief, and we did not test other phrasings. Astra's 733 findings in four waves had a minimum of 0.90 and a mean of 0.992. Every one of its 129 top-severity findings was stated at 0.97 or higher. Argon never stated more than 0.99, and 38 of its 116 window 1 findings and 66 of 216 later ones fell below 0.95. A scale this compressed leaves almost no room to rank anything, and the single finding we refuted from the monitoring wave sat at the top of it: priority P1, confidence 1.00.

The Opus pattern is different in kind: the 8 November review used 0.12 to 0.90 with a mean of 0.493, and only 3 of 193 findings were at or above 0.90. That spread tracked verification, weakly. Of the 68 review-unit P2 findings that went to a verifier (the other 9 verdicts are audit-derived and carry no reviewer confidence), the 52 confirmed ones had a mean stated confidence of 0.574 (range 0.30 to 0.85), the 14 demoted to P3 had 0.55 (0.40 to 0.85), and the 2 refuted had 0.40 (0.35 and 0.45). A gap of 0.17 rests on two refuted findings and is not statistically meaningful. Our reading is narrow: a varying confidence has room to carry information, while one held at 0.95 to 1.00 has almost none. We did not test Argon's or Astra's against outcomes.

Priority scales also differ between vendors.

Reviewer	Findings	P1	P1 share
Gemini 4 Argon, window 1	116	31	26.7%
GPT-6 Astra, monitoring repo	34	28	82.4%
GPT-6.1 Sol, monitoring repo	28	21	75.0%
GPT-6 Astra, four later waves	733	129	17.6%
Claude Opus 5.5, window 1	25	1	4.0%
Claude Opus, 8 Nov	193	5	2.6%

Table 4. Share of findings each reviewer stated as top severity. The two monitoring-repo rows are the same eight units reviewed by two models.

A P1 from Opus on 8 November meant something rarer than a P1 from Astra, and four of those five were demoted on verification. Priorities cannot be pooled across vendors.

5. What Was Real, and What the Reproducers Showed

On the monitoring repository the orchestrator checked all 29 distinct issues by reading the code and, where cheap, running a reproducer. Twenty-eight were real and one was refuted. Twenty-five of the 28 were fixed in the first fix release; 3 were held pending a decision about which copy of the configuration is authoritative. Of the 28, the report marks 24 confirmed and 4 plausible (consistent with the quoted code, not reproduced).

The refuted finding is the instructive one. It was a P1 at confidence 1.00 whose reproducer exercised a code path that production does not use: it parsed a string where production parses a stream. The reproducer genuinely failed in the way the reviewer described, but on a path production never takes. Rerun on the real path, no value leaked.

For the Anthropic review of 8 November we have a fuller record. Its 5 P1 findings (all in the scripts repository) were each verified independently; all 5 were real, 1 stayed P1 and 4 were demoted to P2. Because the reviewers ranked cautiously, we verified all 78 P2 findings: 69 from review units and 9 derived from a post-merge audit (Section 6).

Verdict on P2 findings	Count
Confirmed	61 (52 review-unit, 9 audit-derived)
Demoted to P3	14
Refuted	2
Confirmed by a second verifier on a stronger model	1

Table 5. Outcomes of P2 verification on 8 November. The first three rows are the 77 verdicts; the first verifier declined the fourth finding and a second verifier on a stronger model confirmed it, making 62 confirmed P2.

With the 5 P1 findings, 67 were confirmed for fixing. The two refutations had a sound quote and mechanism, but the harm did not occur given how the module is used.

We did not score reproducers systematically, so there is no rate, but several 8 November cases had a sound quote and mechanism and a reproducer or reachability claim that did not hold. Our rule: read the reproducer, do not trust its verdict.

We have no per-finding verdict record for the Google windows, so we report no real rate for them.

6. Fixes Carry Defects Too, and a Separate Check Finds Them

The 8 November run also audited 203 earlier bug-hunt changes in the fleet-console repository: 174 were sound, 26 incomplete (the fix covered one code path and left a sibling with the same defect), 2 had weak tests and 1 had been reverted or overwritten. By test quality, 148 were behavioural, 54 were source-regex tests (searching the source text instead of running it), and 1 would have passed without the fix. The 26 incomplete audits became findings (9 at P2, 17 at P3), so the run summary's headline count is 219 (5 P1, 78 P2, 136 P3) against the review units' own 193 (5, 69, 119).

The 67 confirmed findings produced 63 fix changes, 21 in the console repository and 42 in the scripts repository, each with a failing-first test.

Outcome of first independent check	Count	Share
Passed	51	81.0%

Outcome of first independent check	Count	Share
Failed on a stale version bump	9	14.3%
Failed on fix correctness	3	4.8%

Table 6. First independent red/green check of the 63 fix changes. Per the run summary, after one repair and a recheck all 63 passed. One of the version failures also broke an existing test and is counted once, in the version row.

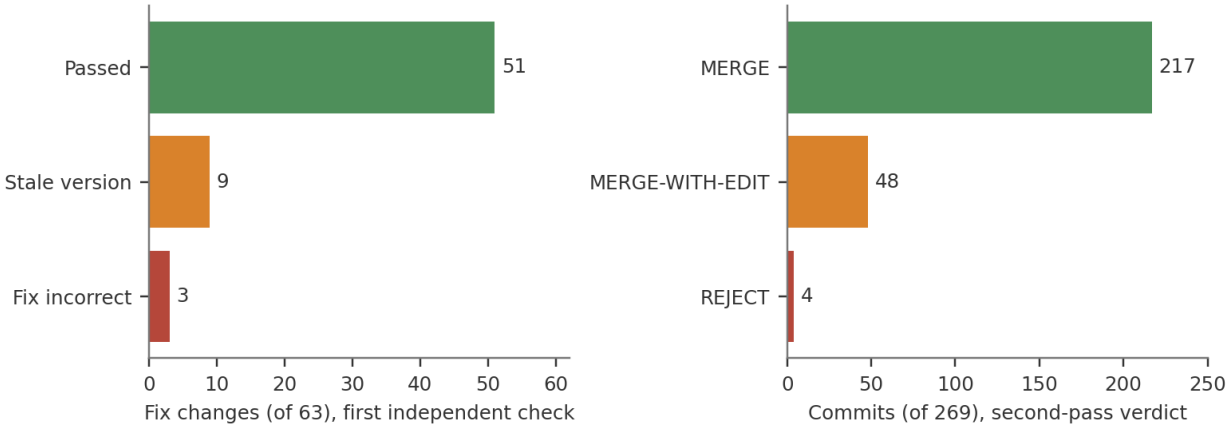


Figure 2. Left: first independent check of 63 fix changes from the Anthropic wave. Right: second-pass verdicts on 269 commits from the xAI wave. Values are in Table 6 and Table 7.

The version failures were a process artefact: other sessions moved the main branch during the run, so a version number assigned at fix time went stale. The three code failures are the point. One was a redaction fix that failed open on one quoting form. One was a guard fix that still let equivalent syntax variants through. One was a spend-guard fix that failed open and left a spend unrecorded, passing silently. A separate check found all three, and the fixes had passed their authors' own tests.

The xAI wave worked differently: 49 code units each found and fixed defects on a branch, a second Grok pass verified each commit, and a third made fix-up edits.

Verdict	Count	Share
MERGE	217	80.7%
MERGE-WITH-EDIT	48	17.8%
REJECT	4	1.5%

Table 7. Verifier verdicts on 269 distinct commits from the 48 verifier units.

All four rejections recorded the defect as not real or not reachable. Of the 217 MERGE verdicts, 214 recorded a real defect with a test failing before and passing after; of 48 MERGE-WITH-EDIT, 46.

We saw three patterns of fixes writing the next round's defects, two of them from single episodes, and one further failure that is not a defect at all.

First, a fix narrows what a guard accepts, and the next input shape slips through. The 8 November guard fix used a pattern that fails open when it misses, although the finding had asked for a fail-closed design, and the check found two equivalent forms the pattern did not cover. The redaction fix introduced a new fail-open, and so did the spend-guard fix. That is three fail-opens in 63 fixes: a count, not a rate.

Second, narrow tests miss suite regressions. On 3 November four guard fixes went into the scripts repository. Per-file runs were green, but the full CI run failed with 61 failed tests and 44

errors, because the new rules blocked legitimate repository tests and one test deleted a directory the tooling needs. We reworked them.

Third, fix passes spawn follow-up fix passes. In the xAI wave's landing logs, 268 distinct commits landed and 57 (21 percent) carry a follow-up in the subject: 35 "follow-up to" an earlier commit, 18 "review follow-up" and 4 "CI follow-up". The last two follow a review or CI result, not necessarily an earlier commit, and some are the planned third pass after a MERGE-WITH-EDIT verdict, so 21 percent is an upper bound on the defects the fixes created.

Separately, merging is not shipping. On 2 November a sweep found six product repositories where fix releases from earlier waves had merged with green CI and never been built or deployed. All six were built and rolled that day. We took that from a session summary and did not recheck it.

7. What These Numbers Will Not Carry

Every result is a single run on different code with a different brief. Sample sizes are 62, 733, 141, 216 and 193 findings, 2 refuted among 77 P2 verdicts, and 1 refuted among 29 monitoring issues. We ran no significance tests; none would mean much at these sizes.

The monitoring-repository classification was one agent's hand check, and elsewhere the agents judging a finding also fixed it. Precision and recall per vendor are unknown, and we did not run Anthropic on code OpenAI had covered, nor the reverse. Wall-clock times are not comparable across vendors.

The orchestrator that classified the monitoring-repository findings, and the agents that verified and fix-checked the 8 November findings, are Anthropic models, so that wave's comparison with the others is not independent of the vendor. We also do not know whether the Opus spread of low-to-mid confidences reflects better calibration or a different reading of the brief.

The claims that survive are these: quote integrity (counted from the retained findings), the stated-confidence distributions, the first-check outcomes for the Anthropic wave, and the verdict counts for the xAI wave (all counted from raw files and not from run summaries). The claims that do not survive are any ranking of vendors by finding quality, and any statement of the form "this vendor's findings are N percent real".

8. What We Changed

Every wave brief still demands a quote, a scenario and a reproducer, but we no longer treat a verbatim pass as evidence that a finding is real. Serious findings from every vendor now get an independent verifier before a fix is written; for the Anthropic wave we extended that to all P2 findings. We do not use stated confidence for triage on OpenAI or Google output; we rank by priority and by whether a fix agent can write a failing test.

The whole repository suite runs before any guard fix merges. Versions for a long fix wave are now assigned at land time, and a wave that merges releases owns their deploy.

9. Methods and Artefacts

The harness used per-unit prompts with a fixed JSON schema, a scorer that checks the quote, and a supervisor that dispatches units in parallel and stops at the allowance wall. The agent CLI for the Google and Anthropic windows ran in a read-only, credential-free sandbox. Verifiers were separate read-only agents. The fix checker copies the worktree, resets changed non-test sources to the base, and runs the new test on both. Statistics are counts and means from the retained JSON. We retain review outputs, merged findings, run logs, verifier verdicts, fix results

and landing logs.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is PT-R-2026-071. The measurements were taken between 26 October and 8 November 2026; the raw artefacts are retained.