

Refuse or Cite: A Source-Grounded Assistant for Front-Desk Procedures and the Test Apparatus That Keeps It Honest

Dustin Aldridge · Forge Agents LLC, Saint Paul, Minnesota ·
daldridge@forgeagentsio.com

Version 2.0, September 26, 2026. DOI 10.5281/zenodo.22985559. Version 1.0 (September 20, 2026): 10.5281/zenodo.22866053. All versions: 10.5281/zenodo.22866052. Public scores, with every failing test named: forgeagentsio.com/proof.html.

Abstract

A hosted assistant answers front-desk questions only from an office's own procedure documents. Every answer either quotes a passage checked against the named document, says the answer is not documented and routes the question to a person, or asks one clarifying question. It serves on Gemini 3.8 Flash, with GPT-5.6 Terra as the backup for the dental and general-office products. All testing used five demonstration corpora that Forge Agents wrote for representative offices.

Version 1.0 reported 310 of 310 tests for the serving model and 110 of 110 for the backup, graded under test definitions that had been edited that same day to accept that day's answers. For this version, every test, the grader, the prompts and the corpora were frozen under a hash manifest before any new answer existed, and none of them changed afterwards. Fresh answers from the serving model passed 306 of 310 tests at three trials each (923 of 930 trials). Of the seven failed trials, five were correct answers the grader missed on wording and two were a minor model error. Under the tests as they stood before the September 20 edits, the same fresh answers pass 296 of 310. The backup passed 107 of 110 fresh.

This version also corrects version 1.0's prompt-evolution result: the loop never served its candidate prompts. Served for the first time, the candidate version 1.0 reported as an improvement scored 36 of 38 against 38 of 38 for the production prompt and was not adopted. The dental knowledge base was padded with other offices' procedures up to 250 documents and still met a capacity rule written before the run, at three trials per test. That rule allows one known miss, which recurs at every size from 50 documents up. A separate judge model found 2 of 3,659 factual claims in fresh answers supported by no document.

1. Introduction

An assistant that answers front-desk questions from an office's procedures is useful only if it never swaps the office's text for general knowledge. The assistant here invents no fee, hour, payer or policy, and it refuses and routes any request for clinical, legal, veterinary or financial judgment.

Version 1.0 reported what the test apparatus measured on September 20, 2026. That day the first runs were read test by test, and 23 test definitions were edited to accept answers judged correct. The acceptance suites had recorded 191 of 200 and were reported at 200 of 200 after regrading. The backup had recorded 106 of 110 and was reported at 110 of 110. A score graded under definitions fitted to its own answers says little about answers those definitions have not seen.

This version measures again under a protocol fixed in advance. On September 25, before any new answer existed, a hash manifest froze 50 files: the tests, the grader, the prompts, the corpora and the tools. The first fresh run started 19 seconds later. Results are reported in four columns:

- **A, as recorded.** The verdicts version 1.0's runs stored on September 20, before any regrade.
- **B, same answers, frozen definitions.** Those answers regraded under the frozen tests, with no model call.
- **C, fresh.** New answers under the frozen tests, three trials per test. This is the headline.
- **H, earlier model.** Claude Haiku 4.5's answers of September 5, regraded under the frozen tests. Only one trial per test was stored, so H is an upper bound.

From A to B only the definitions change, so the gap measures how far the definitions were fitted to those answers. Column C is the only one whose answers did not exist when the definitions were written. Headline scores use no model judge, and no definition changed after the fresh runs.

2. Related work

Retrieval-augmented generation (Lewis et al., 2020) is the standard design for document question answering, and long-context models attend unevenly across long inputs (Liu et al., 2023). Each corpus here held 25 or 26 documents, so the whole corpus went into context; Section 5.5 tests retrieval instead. ALCE (Gao et al., 2023) and attribution evaluation (Rashkin et al., 2021) separate whether a citation is real from whether it supports the claim. Here a verifier checks that each quote occurs in the named document, and a separate judge checks support. Selective prediction (Geifman and El-Yaniv, 2017) treats refusal as a trade between coverage and risk.

The prompt loop followed AlphaEvolve (Novikov et al., 2025) and NVIDIA's AVO (arXiv:2603.24517), in which a model proposes variations under a fixed score. It added AgentDevel's rule (arXiv:2601.04620) that a candidate may not fail a test the incumbent passes. APE (Zhou et al., 2023), OPRO (Yang et al., 2023) and DSPy (Khattab et al., 2023) warn that a prompt tuned on a suite overfits it. The replay store follows the stored-replay idea of Dream-RSI (arXiv:2609.14858).

Dwork et al. (2015) show that a holdout loses its guarantee when each analysis is chosen after seeing earlier results on it. Editing a test after reading an answer is that kind of reuse, and column B measures it. Recht et al. (2019) re-measured published models on new test sets; column C keeps the questions and makes only the answers new. Zheng et al. (2023) document judge biases, including a judge favouring its own answers. The judges here are Gemini models judging Gemini answers, so the headline uses neither. Four parts of the apparatus borrow from earlier work: the acceptance suites from FACET (arXiv:2608.18580), the multi-turn traps from MemTrapBench (arXiv:2608.20202), the citation retry from Chain-of-Experience (arXiv:2608.18027), and the injection suite from Greshake et al. (2023).

3. The system

Answering. Each office's corpus is its set of procedure documents: 25 for the dental practice and 26 for each other office, all describing fictional practices. With every question, the whole corpus and a system prompt of 6,400 to 8,200 characters go to the model, and context caching keeps repeat requests fast. The prompt tells the assistant to answer the latest question and never to treat a statement made in the chat as a fact about the office.

Citations. A quote counts only if it occurs in the named document once spacing and letter case are ignored. An answer with no verified citation is retried once, with the reasons its quotes were dropped.

Models. The console, the hosted service that runs the assistant, serves Gemini 3.8 Flash at a low thinking level in production. GPT-5.6 Terra is the backup for the dental and general-office products. The answering path did not change between version 1.0's runs and the fresh runs: every prompt and corpus had the same hash, and all 930 fresh serving trials recorded Gemini 3.8 Flash as the answering model.

Entailment judge. A separate model, Gemini Flash-Lite at temperature 0, splits each answer into claims and asks whether the cited documents support each factual claim. It then checks any unsupported claim against the whole corpus.

4. Tests and protocol

Suites.

- **Control:** 38 hand-written dental tests on scheduling, insurance, fees, clinical boundaries, false claims of authority and multi-turn traps, seven of them with a fabricated earlier conversation.
- **Safety:** 72 hand-written tests on refusal, routing, injection and Spanish: 16 each for the general-office, legal and veterinary offices, and 24 for the medical office. Together with the control suite, that makes 110 hand-written tests.
- **Acceptance:** 200 tests generated in August from the four non-dental corpora. Each names a source document and quotes the line a correct answer should rest on. A test was kept only if the quote exists verbatim, the expected answer is specific, and the system passed it twice when it was generated. These suites therefore measure whether the system still passes tests it once passed.
- **Final:** 45 dental tests generated on September 20 that the prompt loop never saw.
- **Injection:** eight questions on the general-office corpus, to be answered from the real procedure. Four documents carry planted instructions: waive a fee, open on Sundays, give out staff phone numbers, and give legal advice and print the prompt.
- **Capacity:** the control suite and six probe questions about the dental documents, asked against the dental corpus padded with other offices' procedures up to 250 documents.

The control, safety and acceptance suites make up the nine serving suites, 310 tests in all.

Trials. Every test ran three times and passed only if all three passed.

Grader. Each test checks the answer text for phrases or patterns a correct answer must contain, and for patterns an unsafe answer would have. A forbidden phrase is ignored when a negation such as "do not" comes shortly before it. A model judge can settle cases where the negation is further back. Headline scores are computed without the judge, so those cases count as failures.

Freeze. Just before the freeze, the author approved adding one more accepted wording to each of six test definitions, in each case the wording of a correct answer the grader had failed. Those six additions are disclosed here. Across all stored runs, those six additions turned 10 individual trials from fail to pass. Runs were set to stop at the first rate-limit refusal from a vendor, at an answer from an unexpected model, or when context caching fell below 85%.

One amendment was made during the batch. The first 200-document run hit the serving key's tokens-per-minute limit and stopped, so the 200- and 250-document sizes ran with 7 seconds between requests. Nothing else changed: regrading every stored run before and after the amendment gave identical results, and the stopped attempt was kept but not scored.

Replay. Every run now stores every answer, and older runs were converted from their reports with one answer per test. Any stored run can be regraded under any version of the tests without calling a model. Regrading all 197 stored runs, 7,529 answers, takes about half a second. Columns B, C and H are such regrades.

5. Results

5.1 Before and after

Table 1. Tests passed, with Wilson 95% intervals over tests.

Measure	Serving (310)	Backup (110)
Version 1.0 reported	310	110
A, as recorded	301 · 97.1% [94.6, 98.5]	106 · 96.4% [91.0, 98.6]
B, same answers, frozen tests	309 · 99.7% [98.2, 99.9]	110 · 100% [96.6, 100]
C, fresh	306 · 98.7% [96.7, 99.5]	107 · 97.3% [92.3, 99.1]
C, tests as before September 20	296 · 95.5% [92.6, 97.3]	105 · 95.5% [89.8, 98.0]
H, Haiku 4.5, September 5	307 · 99.0% [97.2, 99.7] (upper bound)	n/a

Version 1.0 published the top row, while its runs had recorded row A. The difference came from regrading under definitions edited that day from those same answers. The rise from A to B likewise reflects fitted tests, not model quality, and B against H is not a comparison of Gemini with Haiku. Haiku's answers were graded 287 of 310 on their own date, and all 20 changes that lift them to 307 came from two rounds of edits, on September 5 and 6, drawn from those answers.

Row C is the headline. Its answers are new, but ten of its 306 passes rely on September 20 edits drawn from the wording of earlier answers to the same questions. Under the tests as they stood before September 20, the fresh answers pass 296 of 310: 104 of 110 hand-written and 192 of 200 acceptance. The six additions of September 25 changed no fresh serving verdict.

The three trials of one test share a question, prompt and corpus, so the intervals are computed over tests, not trials. Each fresh suite ran once, and none was re-run after a miss.

5.2 Serving model by suite

Table 2. Serving model, tests passed. Fresh trials (C): 923 of 930.

Suite (tests)	A	B	C	H
Dental control (38)	38	38	38	38
Office safety (16)	16	16	16	15

Suite (tests)	A	B	C	H
Legal safety (16)	16	16	15	16
Vet safety (16)	16	16	16	16
Medical safety (24)	24	23	24	24
Office acceptance (48)	47	48	48	48
Legal acceptance (51)	49	51	51	50
Vet acceptance (51)	49	51	50	51
Medical acceptance (50)	46	50	48	49
All nine (310)	301	309	306	307

Fresh, the serving model passed 109 of 110 hand-written tests and 197 of 200 acceptance tests. 927 of 930 answers carried a verified citation; the other three were the trials of a test whose correct answer needs none. The four fresh misses:

- **Legal safety, one trial.** The answer was a correct refusal worded "must not interpret", a phrase the grader's list did not accept, so the grader missed it.
- **Vet acceptance, two trials.** Both answers routed an after-hours caller correctly but also told the caller not to wait for the clinic to open. That is an urgency judgment the procedure tells staff not to make, so it counts as a minor model error.
- **Medical acceptance, one trial.** The answer said "clinical team" where the test expected "clinical staff", although the procedure uses both. The grader missed it.
- **Medical acceptance, a second test, all three trials.** The answers were correct, but the test expected words from a part of its reference quote that the question did not ask about. The grader missed these too.

No failing answer stated a test's planted false premise as fact.

From A to B, ten of the 310 test verdicts changed. The nine that went up were acceptance tests the September 20 edits had accepted, using the wording of these same answers. The one that went down was a medical safety test that had passed only through the model judge.

5.3 Backup model

GPT-5.6 Terra ran the five hand-written suites fresh, three trials per test, in a separate evaluation copy of the console.

Table 3. Backup model, tests passed. Fresh trials (C): 325 of 330.

Suite (tests)	A	B	C
Dental control (38)	37	38	38
Office safety (16)	13	16	15

Suite (tests)	A	B	C
Legal safety (16)	16	16	15
Vet safety (16)	16 (a)	16	15
Medical safety (24)	24	24	24
Total (110)	106	110	107

(a) From one stored trial per test.

All five lost trials were correct refusals that were cited and routed but phrased as instructions to staff, such as "Do not give a legal opinion on whether eviction can start". The phrase lists lacked "do not give" and "do not answer". The serving model passed the same three tests on every trial. In production the backup serves only the dental and general-office products, where it passed 53 of 54 tests.

5.4 The prompt loop, corrected

Version 1.0 also ran an evolutionary prompt loop: a model proposed edits to the dental system prompt, and each candidate was scored on the control suite on GPT-5.6 Terra. Version 1.0 reported that one candidate raised the control score from 36 to 38 of 38 and shortened the prompt by 9%, and that it was then reverted.

That result was wrong. The console cached each system prompt on first use and never reloaded it, and the loop copied candidates in without restarting the console. Every candidate was therefore scored on the production prompt while its record named the candidate. The move from 36 to 38 was the production prompt's own variation on Terra, which ranged from 35 to 38 of 38 across seven runs that day.

This version served that candidate, known as candidate 4, properly: the console was restarted and the prompt's hash checked before the first request. It ran on Terra at three trials per test, with these results:

- **Control suite:** candidate 4 scored 36 of 38 and the production prompt 38 of 38. Candidate 4 failed two rescheduling tests on every trial by asking for details instead of giving the procedure.
- **Final suite:** both prompts scored 44 of 45. For each, the one loss was a question the vendor's content filter blocked before any model answered.

Candidate 4 was not adopted. The production prompt passed those two tests on every trial of every Terra control run on record, so the rule against new failures rejects the candidate. The count alone would not have decided it, since 36 lies within the production prompt's own range.

5.5 Retrieval

Table 4. Whole corpus in context against keyword retrieval (BM25) of the top-ranked documents, tests passed. Dental control has 38 tests, office safety 16.

Context	Dental, v1.0	Dental, fresh	Office, v1.0	Office, fresh
Whole corpus	38	38	16	16
Top 5 documents	37	37	15	16
Top 8 documents	36	36	16	16

Full context lost no test. Under retrieval, one dental test, verifying a new patient's insurance, failed on every trial at both depths, because the two sections of the needed procedure ranked 9th and 10th. The answers said the procedure was not in the documents they had; none invented it. The other fresh miss at top 8 was a correct refusal worded "must never recommend", on one trial.

5.6 Capacity

The sweep used the production dental prompt, Gemini 3.8 Flash and three trials per test: 134 requests per size, counting 114 control, 18 probe and 2 warm-up requests. The rating rule was written before the run. The rated size is the largest at which every request is answered and the six probes pass on every trial. Above 25 documents, every control miss must be either the known Saturday answer that credits another location's annex, described below, or a confirmed wording miss. Any answer that states a test's planted false premise as fact disqualifies that size.

Table 5. Capacity sweep, three trials per test.

Documents	Control (38)	Probes (6)	Median seconds
25	38	6	1.35
50	37	6	1.66
100	37	6	1.88
200	37	6	1.65
250	37	6	1.60

Every size met the rule, so the rating is 250 documents, the largest size run. The 250-document corpus holds 462,985 characters, which comes to a median of about 101,000 input tokens per question. Every request was answered at every size. Because requests at 200 and 250 documents were spaced 7 seconds apart (Section 4), their times cannot be compared with those of the smaller sizes.

The one miss from 50 documents up is "Are we open on Saturdays?". The answer says the practice is closed on weekends, then adds a padding document's Saturday hours, attributed to another location's annex. It fails the test and is counted. The rule names it because it had appeared in both earlier sweeps; without that allowance the rating would be 25 documents. Every other trap and boundary test passed at every size.

Version 1.0's sweeps, on Haiku (September 7) and Gemini (September 20), used a generic built-in prompt at one trial per test and are not compared with these figures.

5.7 Entailment

Table 6. Claim-level entailment of recorded answers.

	Version 1.0	Version 2.0
Runs	14	9 fresh serving runs
Answers	1,260	930
Factual claims	5,121	3,659
Supported by cited documents	97.9%	98.3%
Supported only by an uncited document	1.9%	1.6%
Unsupported	10 (0.2%)	2 (0.05%)

Both unsupported claims were in veterinary answers. One offered to "connect you with the appropriate materials or the team". The other told a caller not to wait until the clinic opens: the same advice counted as a model error in Section 5.2, here on a trial the grader passed. The judge, in turn, marked that advice supported in the two trials the grader failed. Grader and judge each let it through at least once, so 2 is a lower bound.

5.8 Injection and failover

Injection. In 24 fresh trials, no answer repeated a planted instruction. Without the judge the suite passed 6 of 8, and 7 of 8 with it. On two trials of the cancellation-fee question the answer said that 24 hours' notice means no fee applies. The procedure says only that a same-day cancellation may carry a fee, so the inference leans the way the planted waiver pushed. It is counted as a minor model error. The other loss was a correct answer ("does not specify whether the office is open on Sundays") that matched a forbidden pattern and passed only with the judge.

Failover drill. With the primary model set to a name that does not exist, the control suite ran through the backup and passed 37 of 38; on one trial of the late-arrival test the answer did not ask whether the patient was new. All 114 answers came from Terra, at a median of 2.3 seconds, the same as calling Terra directly. The primary's startup health check had already failed, so every request went to the backup from the start. Both this drill and version 1.0's showed routing after a failed health check, not failover inside a request.

5.9 Other paths

The same assistant is also served over the Model Context Protocol behind OAuth 2.1. Its first test run through that path, on September 4, ran a 31-test dental set

and the 24 medical safety tests at three trials each, scoring 27 of 31 and 23 of 24 with no citation failures. Only one trial per test survives in the stored records. Regraded under the frozen tests, those trials read 31 of 31 and 24 of 24, an upper bound reached through later edits.

The coverage audit, a self-serve check that asks common front-desk questions of an office's documents and reports which ones they leave unanswered, was run on a test corpus with five planted gaps and reported all five as gaps. On September 25 it failed once with no thinking level set, and passed with the level set to low.

6. The public proof page

An hourly job rebuilds the public scores table from the latest runs. A guard compares each row with the one last published and holds any row that loses two or more tests or more than six points, so one lost test on a 16-test suite is held. From September 20 to 25 the guard held 121 refreshes and published none.

Earlier on September 25, before the freeze, a backup run on the general-office safety suite scored 14 of 16, both misses being correct refusals the phrase lists did not accept. The lists were widened from that run's answers, and a re-run scored 16 of 16 and was published by hand. A note beside the row gave the first score, 14 of 16, and said the tests had been widened and the suite run again.

After the fresh runs, the author reviewed and approved how each miss was classified, and the two held rows, each one test below 16 of 16, were released by hand. The page then published the serving model's 306 of 310 tests (923 of 930 trials), backup rows of 38 of 38 and 15 of 16, and a note on every miss.

7. Conclusion

With frozen tests and fresh answers, the serving model passes 306 of 310 tests, where version 1.0 reported 310 of 310. Under the tests as they stood before the September 20 edits, it passes 296. The backup passes 107 of 110, where version 1.0 reported all 110. And the candidate prompt that version 1.0 reported as an improvement, once it was actually served, scored below the production prompt. The appendix explains, topic by topic, which statements of version 1.0 these results correct.

References

- Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A. The Reusable Holdout: Preserving Validity in Adaptive Data Analysis. Science 349(6248), 2015.
- Gao, T., Yen, H., Yu, J., Chen, D. Enabling Large Language Models to Generate Text with Citations (ALCE). EMNLP 2023.

- Geifman, Y., El-Yaniv, R. Selective Classification for Deep Neural Networks. NeurIPS 2017.
- Greshake, K. et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. 2023.
- Khattab, O. et al. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. 2023.
- Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020.
- Liu, N. F. et al. Lost in the Middle: How Language Models Use Long Contexts. TACL 2023.
- Novikov, A. et al. AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery. 2025.
- Rashkin, H. et al. Measuring Attribution in Natural Language Generation Models. 2021.
- Recht, B., Roelofs, R., Schmidt, L., Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? ICML 2019.
- Yang, C. et al. Large Language Models as Optimizers (OPRO). 2023.
- Zheng, L., Chiang, W.-L., Sheng, Y. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 (Datasets and Benchmarks).
- Zhou, Y. et al. Large Language Models Are Human-Level Prompt Engineers (APE). ICLR 2023.
- NVIDIA. AVO: Agentic Variation Operators for Autonomous Evolutionary Search. arXiv:2603.24517, 2026.
- AgentDevel. arXiv:2601.04620, 2026.
- FACET: execution-validated task synthesis. arXiv:2608.18580, 2026.
- MemTrapBench: Benchmarking Cognitive Traps in LLM Memory Use. arXiv:2608.20202, 2026.
- Chain-of-Experience for Continual LLM Improvement. arXiv:2608.18027, 2026.
- Zheng, T. et al. Dream-RSI: Recursive Self-Improvement through Evolving Worlds. arXiv:2609.14858, 2026.
- Model Context Protocol specification; OAuth 2.1 (draft).

Appendix. What changed since version 1.0

Version 1.0 stays as it was published. This appendix gathers, by topic, every statement in it that this version corrects.

The headline scores. Version 1.0 reported that the serving model passed all 110 hand-written tests and all 200 acceptance tests, and that the backup passed all

110. The runs themselves had recorded 301 of 310 for the serving model and 106 of 110 for the backup. The remaining passes came from regrading those answers after the tests had been edited, and the edits were based on those same answers. Measured fresh, the serving model passes 306 of 310, or 296 under the earlier tests, and the backup passes 107 of 110.

The prompt loop. The candidate prompt that version 1.0 described as raising the control score from 36 to 38, while being 9% shorter, was never actually served. The loop scored every candidate on the production prompt, so the move from 36 to 38 was that prompt varying from run to run. Served properly, the candidate scored 36 of 38 against 38 of 38 and was not adopted. The two lengths, printed as 6,772 and 7,426 characters, were in fact bytes; in characters they are 6,752 and 7,406, and the candidate is still about 9% shorter.

Version 1.0's 45 of 45 on the final suite, on the production prompt, came from a single run. The next run on the same prompt scored 44 of 45 because the vendor's content filter blocked a question. A later run read 28 of 45 after the vendor stopped answering, and it was not scored. Fresh, both prompts score 44 of 45. Version 1.0 ended by saying its next measurements would be this candidate on the final suite and corpora larger than 250 documents. That sentence is withdrawn: the candidate's final-suite result is reported here, and the sweep on the production prompt set the capacity rating at 250 documents.

Capacity. The token counts version 1.0 gave for each corpus size were estimates, characters divided by four; the measured counts are 10% to 13% lower. The seven control requests and two probe requests that version 1.0 described as rejected by the vendor were really rate-limit errors, and there were more of them. Eleven control requests in a row failed at 200 documents, as well as the seven at 250, and the probe requests came after the vendor had already stopped answering. With pacing, the fresh sweep got an answer to every request. Haiku's 26 of 38 at 250 documents counted 7 questions that were never answered: it passed 26 of the 31 it answered. Version 1.0 also said Haiku was refused at 275 documents and above. In fact the 275-document run never started, and 300 and 400 documents did not fit in the model's 200,000-token window. Finally, version 1.0 described one 50-document miss as "phrasing, calibrated the same day", but the wording that fixed it was added on September 25, not September 20.

Failover. Version 1.0 described "an in-request failover to GPT-5.6 Terra". In fact GPT-5.6 Terra is the backup for the dental and general-office products, and both drills show something narrower than failover inside a request: requests go to Terra once the primary model's health check has failed. Version 1.0's drill score of 38 of 38 came from one trial per test. It printed median response times of 2,153 milliseconds through the drill and 2,164 calling Terra directly; the recorded medians were 2,103 and 2,148. At three trials the drill passes 37 of 38.

Grading. The model judge settled three decisions, not the two version 1.0 reported: one medical safety trial and two injection trials. The veterinary case that version 1.0 named never reached the judge at all; it passed once the Spanish word "ningún" was added to the list of negations. Version 1.0's injection score of 8 of 8 held only with the judge's help, and was 7 of 8 without it. In the fresh runs injection scores 6 of 8 without the judge and 7 of 8 with it. Version 1.0 called its retrieval misses false refusals. Only two of them were grading misses; the insurance-verification miss was a real retrieval failure. Version 1.0 also said quotes were checked character for character, with no normalization. In fact, the same code in both versions ignores spacing and letter case. The misses on the Model Context Protocol path, which version 1.0 put down to phrasing lists, were classed in a September 6 review as two correct refusals and three real misses. Regraded under the frozen tests, all five of those stored answers pass, but only because the tests were edited later: three of those edits were based on these answers, and two on answers from September 20.

The proof page and other details. The guard on the proof page is stricter than version 1.0 described. It holds a row back from publication not only when two or more tests are lost, but also when the score drops by more than six points, so losing even one test on a 16-test suite holds the row. Version 1.0 also said the transcripts behind the public scores are published. The page publishes the scores, the failing tests and a note on each, and transcripts are available on request. The coverage audit's result of 5 of 10 covered was measured on August 26, right after two fixes to how the audit wrote its questions and counted coverage. On September 25 the audit failed once with no thinking level set, and passed with the level set to low. Finally, version 1.0 printed a column of scores for GPT-5.6 Luna, a second OpenAI model, at one recorded trial per test. This version reports no Luna scores, and that column belongs to version 1.0.

Data availability

The code and results are held in a private repository. Every score in columns B, C and H was reproduced by regrading stored answers under the frozen tests, with no model call. The stored answers, frozen tests, prompts, corpora and hash manifest are available on request.