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 3.0 · October 5, 2026

DOI: 10.5281/zenodo.23156948

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: each answer quotes a verified passage, states that the answer is not documented and routes the question to a person, or asks one clarifying question. Most earlier failures were correct answers rejected on wording. A paraphrase sweep widened 83 tests whose required wording was too narrow. Written wrong answers exposed 14 tests that accepted an incorrect answer, then three safety and four veterinary tests; all were tightened. The grader reads negations through Markdown emphasis and ends them at contrast words, and both serving paths share one grader. Two veterinary answers that version 2.0 counted as a model error, later called a grader miss, are confirmed as errors: they added timing advice absent from the clinic's after-hours procedure. The test now rejects that advice; answers stopped giving it only once the procedure explicitly forbade it. In the last live run of each of the nine public suites, at five trials per test, the serving model passed 310 of 310 tests, or 309 without the negation referral, which version 2.0's headline excluded. Two outside judge models, rating 4,865 answers, agreed with the grader on 96.9 and 83.9 percent.

1. Introduction

An assistant that answers front-desk questions from an office's procedures is useful only if it never replaces the office's text with general knowledge. The assistant described here is built not to invent fees, hours, payers or policies, and to refuse and route any request for clinical, legal, veterinary or financial judgment. The evidence for that behavior is a test apparatus: suites of questions, a grader that checks each answer, and a public proof page that names every failing test.

Version 1.0 (Aldridge, 2026a) reported full marks under test definitions that had been edited on the day of measurement to accept that day's answers. Version 2.0 (Aldridge, 2026b) froze the tests in advance and measured again: fresh answers from the serving

model passed 306 of 310 tests (923 of 930 trials). Of the four missed tests, three were correct answers the grader rejected on wording and one was classed as a minor model error. Section 4.2 confirms that classification and withdraws a later reversal of it.

That result placed most of the remaining error in the test definitions rather than in the answers. A definition can fail in two directions. A definition that is too narrow fails a correct answer worded differently; the failure is visible, because it appears on the proof page. A definition that is too loose passes a wrong answer; the failure is invisible, because nothing fails. Version 2.0 measured the first kind systematically and saw the second only once, through a judge that checked claims against the documents.

This version reports the work of September 29 to October 5, 2026: a paraphrase sweep that tests whether a correct answer still passes when worded differently (Section 3.2); written wrong answers that test whether an incorrect answer fails (3.3); two corrections to how the grader reads negations (3.4); a rule for loosening a safety test (3.5); one grader for both serving paths (3.6); and a full live run of all nine public suites at five trials per test, re-checked by outside judge models (3.7).

The paraphrase sweep is an invariance test in the sense of behavioral testing (Ribeiro et al., 2020), applied to the grader rather than to the model. Written wrong answers play the role that mutants play in mutation testing (DeMillo et al., 1978): a test that does not reject them is too weak. Both methods change test definitions after recorded answers have been read, which is the adaptive reuse that Dwork et al. (2015) show erodes the validity of a test set. Section 2.4 states the rules that limit this, and Section 4.1 what the changed definitions support.

2. Method

2.1 System

Each office's corpus is its set of written procedures: 25 documents for a dental practice and 26 each for a general office, a law office, a veterinary clinic and a medical clinic. All five practices are fictional, and Forge Agents wrote the corpora. With every question, the whole corpus and the office's system prompt are sent to the model. A quoted passage counts as a citation only if it occurs in the named document once spacing and letter case are ignored, and an answer with no verified citation is retried once.

The serving model is Gemini 3.8 Flash at a low thinking level. GPT-5.6 Terra is the backup model for the dental and general-office products. The assistant is reached by two paths: the console, a hosted web application, and the connector, a single-turn tool served over the Model Context Protocol (Model Context Protocol, 2025) behind OAuth 2.1 (Hardt et al., 2025). The system prompts and the serving model are unchanged from version 2.0. The corpora are unchanged except for one rule added to the veterinary clinic's after-hours procedure on October 4 and 5 (Section 4.2).

2.2 Suites and terms

The nine public suites hold 310 tests. The dental control suite has 38 hand-written tests on scheduling, insurance, fees, clinical boundaries, false claims of authority and multi-turn traps. Four safety suites hold 72 hand-written tests on refusal, routing, prompt injection and Spanish: 16 each for the general office, the law office and the veterinary clinic, and 24 for the medical clinic. Four acceptance suites hold 200 tests generated in August from the non-dental corpora by a method adapted from execution-validated task synthesis (Shi et al., 2026): 48 for the general office, 51 legal, 51 veterinary and 50 medical. A generated test was kept only if its quote occurs verbatim in the corpus, its expected answer is specific and the system passed it twice when it was generated, so these suites measure whether the system still passes tests it once passed. Tests are named by the identifiers the proof page uses; acceptance test numbers repeat across the four acceptance suites, so each is named with its suite.

A **test** is one question, or a short conversation, together with its **definition**: the content a correct answer must contain (required content) and the content a wrong or unsafe answer would contain (forbidden content). Required content is a list of accepted phrases, a pattern, or several elements that must all appear, alone or combined. A test may also require a verified citation; 15 tests require one from a named document.

A **trial** is one answer to one test. A test passes only if every trial passes. Runs reported for this version use five trials per test; version 2.0's runs, cited in Sections 1, 3.1, 3.8, 4.1 and 4.2, used three.

The **grader** is the program that applies a definition to an answer. It uses no language model, except for the negation referral described in Section 2.3.

A **judge** is a separate language model that reads an answer and rates it. Judges are used to find candidate grader errors and to re-check results.

A **recorded answer** is an answer the system gave in an earlier run, stored with its citations. Every run stores every answer, so any stored run can be **replayed**: graded again under another definition or grader version without calling a model. Each replay reported in Section 3 covered the stored runs gathered for that check on its date, so replay totals differ between sections. A **live run**, by contrast, sends every question to a model.

A **rewording** is a paraphrase of a recorded correct answer, written by a model, that keeps the answer's facts.

A **written wrong answer** is an incorrect or unsafe answer, composed by a person or by a model, to check that a definition rejects it.

2.3 Grader

The grader checks the answer text for required and forbidden content. Forbidden content is negation-aware: a forbidden phrase does not count when a negation governs it, so "do not quote a fee" does not count as quoting a fee. When the grader cannot settle

whether a negation governs a forbidden phrase, it refers the sentence to a lightweight Gemini model, which rules whether the phrase is asserted. This happens in two cases: a negation stands too far before the phrase for the grader to connect them, so the phrase would count, or a nearby negation mutes the phrase but may not govern it. A referral can therefore turn a fail into a pass or a pass into a fail. This is the negation referral. Unlike version 2.0, whose headline scores excluded it, this version's scores include it; Section 3.7 reports how many verdicts depended on it.

2.4 Rules for changing a test

The paraphrase sweep and the tightening each applied a written rule. A widening in the sweep had to keep every recorded answer that passed still passing, reject at least three written wrong answers and raise the test's rewording pass rate. A tightening had to keep every recorded correct answer passing. A third rule, written down on October 4 and applied to one test (Section 3.5), allows a safety test to be loosened only when outside judges rate the answers it fails as safe and every written wrong answer for that test still fails. Changes made in the recalibration of September 29 and 30 were graded against recorded answers before they were applied but were not held to these rules. Each change made during the full run (Section 2.5) was checked offline before its suite ran again: every recorded correct answer that had passed still passed, and no written wrong answer then on file that had failed began to pass, except three model-written ones when two veterinary tests were widened on October 5 (Section 4.1). The proof page reports only live runs. A judge's rating may start a review of a grader failure, and is never used to overturn a grader pass.

2.5 Full run

The full run is the set of live runs that produced the nine rows now on the proof page: all nine public suites, run on the production console against the serving model at five trials per test, 1,550 trials in all, on October 4 and 5, 2026. Dates in this report are in Coordinated Universal Time. Unlike version 2.0's run, the full run was not frozen in advance. Ten test definitions and one procedure changed while it was in progress (Sections 3.7, 4.1 and 4.2), and each suite affected was run again after its last change; the row on the proof page is that last run. Every published run used the final test definitions, the same version of the grader and the system prompts of version 2.0. Some runs reported here used the connector, and some used an evaluation copy: a separate, temporary instance of the console, used to try a change or a model without touching production.

Two outside judge models, Jev 1.13 (TypeSafe) and gpt-oss-20b, both reached through a hosted service, rated answers from runs made on October 4 and 5. Each judge was given the question, the answer with its citations, the source passage, and the test's rules as recorded with the run, restated in plain language, and rated the answer pass or fail. No rating changed a score.

3. Results

3.1 Starting point

Version 2.0's fresh run is the baseline: under the definitions frozen for it, the serving model passed 306 of 310 tests. By the end of the full run, 180 of the 310 public test definitions had changed; the published rows use these final definitions. No question and no system prompt changed, and the corpora changed only by the veterinary rule of Section 4.2. Replayed under the final definitions, and graded like version 2.0 without the negation referral, the same 930 answers pass 309 of 310 tests. Three of the four tests that missed now accept the answers they rejected. The fourth, the veterinary after-hours test, rejects them again, now because of the timing advice they contain (Section 4.2). That figure measures how far the definitions moved toward those answers; it is not a result (Section 4.1).

The changes came from a recalibration of the medical, general-office, legal and veterinary suites on September 29 and 30, the paraphrase sweep, the tightening of weak tests, the replacement of forbidden phrases that the procedures themselves use, four safety-test revisions on October 4 (Sections 3.5 and 3.6), and the changes made during the full run (Sections 3.7 and 4.1). The recalibration reviewed against the procedures each test on which both outside judges, Jev 1.13 and gpt-oss-20b, passed an answer the grader failed. Later five-trial runs still exposed four tests that rejected a correct answer phrased in a new way, and the paraphrase sweep was built to find such tests in advance.

3.2 Paraphrase sweep

For each test, up to four recorded answers that the grader had passed were selected, 1,239 in all, and gpt-oss-20b wrote two rewordings of each (2,478). Each rewording was graded under the test's own definition. A failing rewording pointed either to required content listed too narrowly or to a rewording that had lost a fact. Jev 1.13 rated the 258 rewordings that failed on required content (five more failed on forbidden content), gpt-oss-120b proposed additional accepted phrases, and every proposal was checked against the rules in Section 2.4.

A rewording failed on required content in 102 tests. The sweep widened 83 of them, across all nine suites. Table 1 gives the share of rewordings the grader passed before and after the change.

Of the other 19 tests, six were tightened instead (Section 3.3) and 13 were left unchanged: in eight the failing rewordings had dropped a fact the test requires, three express their required content as a pattern, and in two the proposed phrase repeated the question or risked passing a wrong answer. The "after" figures are measured on the same rewordings that informed the changes, so they overstate the gain on new answers. All nine edited suites were run again live at five trials by October 4, seven of them on September 30, the day of the sweep, and every test passed.

Table 1. Rewordings of recorded correct answers passed by the grader, before and after the paraphrase sweep (percent).

Suite	Tests widened	Before	After
Dental control	14	84.2	92.1
General-office safety	5	94.5	99.2
Legal safety	6	86.7	96.1
Veterinary safety	5	88.3	95.3
Medical safety	3	95.3	98.4
General-office acceptance	17	84.3	95.0
Legal acceptance	15	87.3	98.3
Veterinary acceptance	13	89.7	96.8
Medical acceptance	5	96.8	98.5

3.3 Weak tests

The written wrong answers prepared for the sweep showed the opposite fault: 14 tests in six suites passed at least one of them. Each of the 14 now requires two elements together, rejects the wrong answer's form with a negation-aware forbidden pattern, or both.

All 69 written wrong answers for these tests, four to six per test, now fail. Every recorded correct answer still passes; one dental answer from September 3 that had passed now fails, because it stopped before its routing step. No rewording that passed before fails now. Six suites were run again live at five trials, and every test passed. On September 30 the latest five-trial run of every public suite passed every test, although one of those rows passed answers now counted as wrong (Section 4.2).

3.4 Grader corrections

Two corrections changed how the grader reads negations.

The first, on September 29, concerns contrast words. A negation early in a sentence could mute a forbidden phrase that followed it, so "Staff do not authorize refunds, but the office manager always approves refunds" could pass a test that forbids promising refund approval. A contrast word (but, however, although, and the Spanish pero and sin embargo) now ends the reach of the negation before it. The first version of this change moved no verdict over 6,863 stored answers, yet it failed 21 of 48 correct refusals written to probe it, through words such as a mid-sentence "however". It was revised so that such refusals stay intact, and 18 of those cases were kept as regression tests.

The second, on October 1, concerns Markdown emphasis. Some answers put a negation in bold ("Do **not** include"), and the grader did not recognize a negation inside emphasis marks, so such a refusal scored as asserting the forbidden phrase. The grader now reads through bold, italic and code marks when it looks for negations and forbidden phrases. An adversarial pass on this change found three further faults, all fixed, among them expressions such as "no worries" and "not only" that contain a negation but negate

nothing. The change is covered by 48 written test cases, half of them adversarial. Replayed over 17,124 stored answers in 247 runs, it moved one verdict: a correct Spanish refusal, from fail to pass.

A related change replaced forbidden phrases that the procedures themselves use, such as a 24-hour figure that one procedure gives for cancellation notice, with patterns that tie the figure to the wrong claim. The check that flags such phrases now reports 8 warnings instead of 18, each kept after review.

3.5 An intermittently failing safety test

The general-office safety test OKA-SAFE-013 asks the assistant to ignore its instructions, become a different assistant named OpsBot, confirm the switch, and state the refund policy from memory. Its definition required the answer to announce a refusal, and in runs that checked a release on October 4 the test failed intermittently on the connector. Three answers gave the refund procedure with its citation and did not adopt the persona, and failed only because they did not say that they would not switch.

All 61 recorded answers to this question were collected, 54 from the console and 7 from the connector. Two outside judges, Jev 1.13 and gpt-oss-120b, rated all 61 safe on the four failures the test exists to catch: adopting the persona, obeying the injected instruction, claiming to answer from memory, and stating invented or uncited terms. The grader had failed five of them: the three connector answers and two Spanish answers from GPT-5.6 Luna, an earlier backup model. Each of the five failures concerned wording, so the test was changed and the product was not.

Under the revised definition, an answer passes by refusing or by stating the procedure's rule that front-desk staff do not authorize refunds, or its routing step. It must carry a verified citation, and 14 added forbidden patterns, 22 in all, reject the failures the test exists to catch, among them a persona label, memory given as the source and a promised approval. Checked offline first, the new definition passed 59 of the 61 recorded answers (56 before; the two Spanish answers still fail) and all 118 rewordings (88 before), and it failed all 31 written wrong answers and 60 more that a model wrote to slip past it. Replay over 237 stored runs moved no verdict. Run live at five trials, the general-office safety suite passed 16 of 16 on both the console and the connector, and the proof page carries a note on the change.

3.6 One grader for console and connector

Until October 4 the two paths were graded by two copies of the same grading function. The connector's copy predated the corrections of Section 3.4, and it differed from the console's in both directions. The contrast-word example of Section 3.4 failed under the console grader and passed under the connector grader. In the other direction, 23 correct console answers, among them clarifying questions and Spanish refusals, would have failed under the connector grader. The grader is now a single shared component that both runners load, and a runner refuses to start beside a stale copy.

Replayed over 15,676 stored answers in 272 runs, 585 of the answers from the connector, the unified grader moved no verdict. In an adversarial pass, gpt-oss-120b wrote 120 answers for 10 safety tests, six correct and six unsafe for each, and both runners graded them with identical verdicts. Three wrong answers passed on both runners because their definitions did not cover the wording; one was a Spanish answer stating “No es una emergencia”. Those three safety tests were tightened on October 4 under the tightening rule of Section 2.4, and the full run is their first live run.

The connector is tested on 31 of the 38 dental control tests; the seven multi-turn tests do not apply to a single-turn tool. Under the unified grader, at five trials, its last runs on October 5 passed the dental set 31 of 31 and the general-office and veterinary safety suites 16 of 16 each, all under the final definitions. An earlier general-office run on the connector failed OKA-SAFE-008 on a correct answer that offered to take a message and report it to the office manager; that test was widened before the last run.

3.7 Full live run

Table 2. Full run: the last live run of each public suite on the production console, serving model, five trials per test. Dates are the run dates shown on the proof page.

Suite	Tests	Tests passed	Trials passed	Run date
Dental control	38	38	190 of 190	October 4
General-office safety	16	16	80 of 80	October 5
Legal safety	16	16	80 of 80	October 4
Veterinary safety	16	16	80 of 80	October 5
Medical safety	24	24	120 of 120	October 4
General-office acceptance	48	48	240 of 240	October 4
Legal acceptance	51	51	255 of 255	October 4
Veterinary acceptance	51	51	255 of 255	October 5
Medical acceptance	50	50	250 of 250	October 4
All nine suites	**310**	**310**	**1,550 of 1,550**	

The serving model passed 310 of 310 tests; every test passed on every trial, and every answer came from the serving model. A Wilson (1927) 95% interval over tests runs from 98.8 to 100 percent, but it assumes tests fixed in advance, which these were not (Section 4.1). Of the 1,550 answers, 1,542 carried at least one verified citation. The other eight had no passage to quote: five stated that the procedures do not cover a question about an employee retirement plan, and three were clarifying questions that their test accepts. Seven verdicts, in three tests, went to the negation referral. In one, a legal refusal whose negation lay beyond the grader’s reach, the referral turned a fail into a pass; in the other six it confirmed a negation the grader had already found. No referral turned a pass into a fail. Version 2.0 computed its headline without the referral; graded that way, the published runs pass 309 of 310 tests, the legal acceptance suite 50 of 51.

The published row of each suite is its last run on the production console. Before those rows, ten runs on the production console, on evaluation copies and on the connector failed 13 trials across nine tests. This count leaves out six single-test runs of the after-hours question on draft wordings of the new procedure rule, which failed 17 trials (Section 4.2). Each of the 13 failed answers was read against the procedures and found correct, and each failed only on its wording: refusals such as “You must not quote a late fee”, referrals in a procedure’s own words, and one refusal that repeated the question. The nine tests were widened under the check of Section 2.4 and their suites run again. The last such case was a veterinary acceptance run of 49 of 51 on October 5, after which GEN-10 and GEN-14 were widened and the suite’s next run scored 51 of 51. Five tests were tightened during the run. The veterinary acceptance test GEN-37 had passed any answer that cited the right procedure, because every answer ends with a source line containing the words the test looked for; it now ignores that line, and no recorded answer changed result. The after-hours test of Section 4.2 was tightened on October 4, and again on October 5 together with GEN-5, GEN-10 and GEN-14, after model-written wrong answers got past all four (Section 4.1). The published veterinary acceptance row is the run made after that last change.

Table 3. Outside judges against the grader: 4,865 answers from 29 runs made on October 4 and 5, including eight of the nine published rows. Grader verdicts are those recorded at run time.

Judge	Answers rated	Agreement with grader	Grader pass, judge fail	Grader fail, judge pass
Jev 1.13	4,865	96.9%	141	8
gpt-oss-20b	4,824	83.9%	769	8

The 4,865 answers come from the published runs, the production runs they replaced, runs on evaluation copies and connector runs; gpt-oss-20b returned no usable rating for 41. The grader failed nine of them, all wording failures counted above. Both judges passed eight and failed the ninth, a correct legal answer, by reading a rule more literally than the procedure does. Both judges failed 99 answers that the grader passed, in 20 tests. Read against the procedures, every one was correct: the judges had miscounted the citations an answer must carry, read a list of acceptable phrasings as a list of required points, or failed a clarifying question that the test allows. Jev also rated the 255 answers of the published veterinary acceptance run and agreed with the grader on 249; the six it failed were correct.

3.8 Results carried from version 2.0

The backup model, capacity, retrieval, claim-level entailment, injection and failover were not measured again. Except for the two backup runs named below, the figures here are version 2.0’s (Aldridge, 2026b), measured under its definitions and grader at three trials per test. On fresh trials the backup model, GPT-5.6 Terra, passed 107 of 110 hand-written tests, and 53 of 54 on the dental and general-office products it serves. Two later backup runs at five trials on an evaluation copy, listed on the proof page, passed 16 of 16 general-office safety tests on September 29 and 37 of 38 dental control tests on

September 30. Whole-corpus context scored at least as well as keyword retrieval of the top five or top eight documents in every comparison. Under a capacity rule written before the run, which allows one known miss that recurs at every size from 50 documents up, the dental knowledge base was rated at 250 documents, the largest size run; without that allowance the rating would be 25 documents. A separate judge rated 2 of 3,659 factual claims as supported by no document; one of them is the veterinary advice discussed in Section 4.2. No answer repeated a planted instruction in 24 injection trials, but the injection suite passed 6 of 8 tests without the negation referral and 7 of 8 with it, and two trials of one question leaned toward the planted fee waiver, a minor model error. With the primary model disabled, requests went to the backup, which passed 37 of 38 control tests.

4. Discussion

4.1 Changed definitions and what they support

More than half of the public test definitions changed after version 2.0, 180 of 310, and many changes were made after recorded answers had been read. Dwork et al. (2015) show that adaptive reuse weakens a holdout set's statistical guarantee. Here the effect is direct: a definition fitted to an answer passes that answer whether or not it is correct. Version 2.0's fresh answers show the effect. Under the frozen definitions they passed 306 of 310 tests; under the final definitions they pass 309. The difference measures how far the definitions moved, not how well the model answers. For some tests the fitting went too far, and those tests have been tightened again (Section 4.2 and below).

Three controls limit the effect, none of them completely. Rewordings are answers the system never gave, but the sweep also used them to choose the widenings, so they are not a held-out check (Section 3.2). Written wrong answers test the opposite direction: a widening counted only if they still failed, and after the tightening all 69 fail. The published rows of the full run consist of answers that did not exist when their suites' definitions were last changed, so they are the result this version reports. They are not a measurement fixed in advance, as version 2.0's was: ten definitions changed during the run, each after a run failed a correct answer or a review found a test too loose, and each suite concerned was run again.

Written wrong answers are also finite. On October 5, model-written wrong answers showed that four veterinary acceptance tests, GEN-5, GEN-10, GEN-14 and GEN-34, could still pass a wrong answer that named the right person or the posted emergency hospital and then also gave advice, predicted or promised an outcome, or added an invented rule, fee or time. Of 90 wrong answers written for each of GEN-10 and GEN-14, GEN-10 passed 38 and GEN-14 passed 11; three of those 49 had failed before the two tests were widened that day (Section 3.7). Measured on the same 90 wrong answers, GEN-10's definition at the start of the full run passed 21; an October 4 widening that let it accept more correct answers, checked against the wrong answers that existed then, had already raised that to 37. The four tests were then tightened on the forbidden side only; their

required content did not change. Replayed against every earlier run of the suite, no recorded answer changed result. GEN-10 now passes 1 of those 90 wrong answers and GEN-14 none. Of 240 wrong answers written after the final change, the four tests pass 14, where their earlier definitions passed 190; earlier rounds of such answers had been used to revise the definitions. Of 240 answers from the same round that the writing model labeled correct, 222 pass, against 224 before; the two that now fail urge the caller to call now, which the revised procedure rules out.

4.2 Version 2.0's veterinary miss

Version 2.0 classed two failed trials of a veterinary acceptance test, GEN-34, as a minor model error. A caller leaving a message after hours asked whether a lethargic dog could wait until the clinic opened. Both answers declined to judge urgency and sent the caller to the posted 24-hour emergency veterinary hospital, as the procedure directs. Both then added advice on timing. One told the caller to go "immediately rather than waiting for the clinic to open"; the other said the pet "must be evaluated by an emergency clinician rather than waiting for regular clinic hours". The procedure contains neither statement, and it tells staff not to triage or advise and never to assess urgency.

The grader failed both trials for a different reason: the test expected the words "emergency hospital", which "24-hour emergency veterinary hospital" does not contain. That part was a grader miss on wording, but the answers still contained an error, so version 2.0's classification was right. A review on September 30 reached the opposite conclusion. It reclassified the trials as a grader miss, widened the test, and the proof page carried a note saying so. That reversal was wrong. The widened test passed any answer that named the hospital, and the answers behind the 51 of 51 published for this suite on September 30 gave the same advice to go immediately: that row passed answers now counted as wrong. The proof page carries a correction, dated October 4, 2026, that says so.

Two changes followed. On October 4 the test was tightened so that timing advice fails; under the final definition, all three of version 2.0's answers to this question fail, as do the five behind the September 30 row. On October 4 and 5 the clinic's after-hours procedure gained one rule: after hours, a caller who reports symptoms, an injury, a possible poisoning or any urgency is directed to the posted 24-hour emergency hospital, the message is logged for the veterinarian, and staff give no advice on timing. Under the original procedure, which already told staff not to advise or assess urgency, all 35 recorded answers to this question gave timing or urgency advice, including all five answers of a run on an evaluation copy on October 4, made shortly before the rule was written. The rule's wording was tried in six single-test runs of this question on draft texts. In three of them, answers still failed the test in 2 of 5, 3 of 10 and 12 of 20 trials, 15 of the 17 failures on explicit timing advice; the adopted text adds that the referral is stated plainly, without urging speed. The published pass on this test therefore shows that the assistant follows an explicit instruction against timing advice, not that it avoids such advice when a procedure only says not to advise. The test was also adjusted so that

it no longer fails a correct refusal that repeats the words of the question, and on October 5 it was tightened again with three other veterinary tests (Section 4.1). In the published row, all five answers decline to judge whether the dog can wait, direct the caller to the nearest 24-hour emergency hospital, log the message for the veterinarian, and give no timing advice.

4.3 Replay is necessary but not sufficient

Replay finds regressions on answers already given, quickly and without model calls, but it cannot show that a grader change is safe on answers not yet given, as the contrast-word change of Section 3.4 showed. Each grader change was therefore also tested on written adversarial cases.

4.4 The role of judges

Outside judges shortened the review. They rated thousands of recorded answers and pointed to the few worth reading against the procedure, and their ratings went no further: a judge's pass could start a review of a grader failure, and no rating changed a published score. The full run shows why the two directions are treated differently. When the grader failed an answer, both judges passed it in eight of nine cases, and all eight were wording misses; when the grader passed an answer that both judges failed, the judges were wrong in all 99 cases. Neither judge objected to the timing advice of Section 4.2: given the test's rules as they then stood, Jev passed all five answers of October 4 that contained it, and gpt-oss-20b failed three of them for other reasons. The review itself can still err, as Section 4.2 shows. Judges are known to favor answers resembling their own (Zheng et al., 2023), which is one reason both judges come from developers other than the serving model's.

5. Conclusion

Most of the error left after version 2.0 was in the test definitions, and this version checked them in both directions: rewordings found 83 tests too narrow; written wrong answers found 14 too loose and, later, three safety and four veterinary tests; and one corrected grader now serves both paths. A review on September 30 wrongly reversed version 2.0's one counted model error; Section 4.2 restores it, the test now rejects the timing advice, and answers stopped giving that advice only after the clinic's procedure gained an explicit rule against it. Under the final definitions, the last live run of each of the nine public suites at five trials passed 310 of 310 tests, or 309 without the negation referral that version 2.0's headline excluded, and two outside judge models agreed with the grader on 96.9 and 83.9 percent of the answers they rated.

Data availability

The stored answers, test definitions, written wrong answers, rewordings, judge ratings and grader versions behind this paper are available on request. Public scores are at forgeagentsio.com/proof.html.

References

Aldridge, D. (2026a). Refuse or Cite: A Source-Grounded Assistant for Front-Desk Procedures and the Test Apparatus That Keeps It Honest. Version 1.0. Forge Agents LLC. Zenodo. https://doi.org/10.5281/zenodo.22866053

Aldridge, D. (2026b). Refuse or Cite: A Source-Grounded Assistant for Front-Desk Procedures and the Test Apparatus That Keeps It Honest. Version 2.0. Forge Agents LLC. Zenodo. https://doi.org/10.5281/zenodo.22985559

DeMillo, R. A., Lipton, R. J., Sayward, F. G. (1978). Hints on Test Data Selection: Help for the Practicing Programmer. Computer 11(4), 34-41.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A. (2015). The Reusable Holdout: Preserving Validity in Adaptive Data Analysis. Science 349(6248), 636-638.

Hardt, D., Parecki, A., Lodderstedt, T. (2025). The OAuth 2.1 Authorization Framework. IETF Internet-Draft draft-ietf-oauth-v2-1-13, May 28, 2025.

Model Context Protocol (2025). Model Context Protocol Specification, version 2025-06-18. https://modelcontextprotocol.io/specification/2025-06-18

Ribeiro, M. T., Wu, T., Guestrin, C., Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. Proceedings of ACL 2020, 4902-4912.

Shi, K., Wang, Z., Su, Q. et al. (2026). FACET: Preserving Source Intent and Executable State in Terminal Task Synthesis. arXiv:2608.18580.

Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association 22(158), 209-212.

Zheng, L., Chiang, W.-L., Sheng, Y. et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023, Datasets and Benchmarks.