THE CONVENTION IS THE GRADER'S: HOW MODEL PROOF GRADERS PRICE SILENTLY REPAIRED ERRORS

Anonymous authors
Paper under double-blind review

ABSTRACT

The grading rubric bands a small repairable error at 5–6 and says nothing about a false line the proof itself corrects; what it costs is the grader's judgement. We measure that price and propose no grader. Into 24 IMO 2026 proofs, model-judged closed, not human-verified, we plant three single-hunk edits for 96 texts — fatal, harmless (nothing later uses it), or a deleted argument. Falsity is machine-checked; class is a recorded label (8 of 72 disputed). A panel of 5 graders scores the texts, and every decision rule is read under two conventions for the repaired line. Graders notice harmless slips and price them differently: on the 15 harmless texts where both pair graders quote the planted slip, one awards full marks on 2/15 and the other on 14/15, reversing which grader looks most accurate on the strict mark; a retest keeps the reversal. The completeness flag sits between score thresholds and differs from them only within grade 6. An obligation audit whose prompt says not to repair, read as a comparator, orders the graders the same way (12/24 against 24/24 harmless mutants accepted under its rule); its held-out gap to the strict mark is exploratory and unresolved ($p \geq 0.18$; its gain three announced deletions), and a pre-registered replication confirms only its gap to the pass mark. On human-majority-labelled competition proofs, the strict mark and flag err mostly by false rejection. With the audit's no-repair sentences in the grade prompt, the strict mark moves to 6/32 and 13/32 (exploratory).

1 INTRODUCTION

Frontier models produce proofs of olympiad problems faster than anyone can read them, so a grader decides whether a proof is correct, and a pipeline that must decide turns its 0–7 score into a decision by a threshold. The rubric the grader is given fixes the scale and leaves one thing open: it bands a small, straightforwardly repairable error at 5–6, and it is silent on a false line the proof itself corrects a line later — whether that line is an error at all, and whether the proof's own repair moves the mark. A pipeline has to answer that before the score means anything, and when nobody answers it the grader does.

Into 24 proofs of the IMO 2026 problems that each model evaluator which saw them judged closed — none verified line by line by a human — we plant three single-hunk edits, one that falsifies a statement a later line uses, one that falsifies a display nothing downstream reads, one that deletes an argument and keeps its claim, for 96 texts (§2). Falsity is machine-checked by an executable refutation, class is recorded by a quoted dependency certificate a reader checks by eye, and the price a coordinator would put on the edit is neither; we say whose judgement it is. A panel of 5 graders scores the texts, 2 of which also run an obligation audit, and every rule is read under two stated conventions for a repaired line, as written and repair-tolerant (§3); two thirds of the corpus is held out.

The graders fill the rubric's silence differently. On the harmless class — a false display nothing later uses, usually corrected by the next line — a grader that quotes the planted slip awards full marks anyway on 3/20 of the texts where it does so for one grader of the decision pair and on 17/18 for the other; told in the audit prompt not to repair, one accepts 12/24 harmless texts and the other 24/24. In the stored run the difference reverses which grader looks most accurate between the two readings on the strict mark, the completeness flag and the audit alike (3 rules agreeing pairwise on 76–90 of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

96 texts; the audit read on the pair alone; §5). The reversal follows from the design; the size of the difference between two graders under one prompt does not, and it is the finding. The convention is the grader's and survives a re-run; its other prices do not: re-run unchanged, the prompt keeps the difference and the strict mark's reversal, not the flag's, and lifts one grader's strict mark on 13 texts, 12 of them fatal or gap edits, lowering it on none (§4).

None of the fields we read encodes the convention. The completeness flag is nested between two score thresholds — over 469 first-corpus gradings from the 5 graders it is set on 284/284 of the responses graded 7, 0/41 of those graded 5 or below and 27/144 of those graded 6 — so what it adds to the score lies inside grade 6, where for the pair's graders it sorts 2/30 and 2/17 held-out texts, in opposite directions; and among the 34 held-out texts whose lower pair grade is 6 sit 14 fatal, 7 deletion, 10 harmless and 3 original texts, so no threshold on that score separates a repaired line from an unrepaired one (§4). Nor is a pass mark's leniency a detection failure: of the broken texts a pass mark of 5 accepts for the pair, at least one grader quotes inside the planted hunk on 44/45 (23/45 for one grader, 43/45 for the other, by overlap of the quote with the edited span), at a median grade of 6 over the pair's grades on those texts; the response holds the objection, and neither it nor the rubric says what the edit costs.

The one rule that costs a second call, an obligation audit whose prompt says not to repair, we read as a comparator, not a proposal, its dependence first: its gain over the strict mark on the broken axis is three announced deletions (§F). Counted repair-tolerant on the 16 held-out bases (exploratory; §6) the audit accepts fewer broken texts and rejects fewer sound ones than the strict mark, 4/32 and 8/32 against 7/32 and 13/32, on discordant pairs too few to settle ($p \geq 0.18$, §4); as written it is a trade; a pre-registered replication on 48 unseen texts confirms only its gap to the pass mark (§7). With its no-repair sentences in the grade prompt (exploratory) the strict mark moves to 6/32 and 13/32 (§6). On 254 multi-judged proofs with a majority human label (itself a pass mark of 5 by the judge instructions) the pass mark errs on 11, the strict mark on 52 and the flag on 47, mostly false rejections (§8).

Contributions. (i) The convention is the grader's: under one rubric, how much a quoted harmless slip costs differs between graders enough to reverse which looks most accurate (§5). (ii) A probe corpus whose labels say what they establish — falsity by executable refutation, class by a recorded dependency certificate — and what they do not: whole-proof correctness and the price of an edit (§2). (iii) The completeness flag differs from the strict mark only within grade 6, where its few distinctions run opposite ways per grader, and no threshold separates a repaired line from an unrepaired one on that band (§4). (iv) Every rule one response supports, per grader with the pair's retest beside it, the audit among them (exploratory held out; its gain named; a replication confirming only its gap to the pass mark), and the rules read against a majority human label (§§4–8).

2 A Mutant Suite with Executable Falsity Checks

Bases. We take 24 proofs of the six IMO 2026 problems (IMO, 2026) that every model evaluator which saw them judged structurally closed, and "accepted" below means that verdict and nothing more. The evaluators are the ones the base-selection protocol names — GLM-5.3, Kimi K3, Claude Opus 5 and GPT-5.6-Sol under the closure and replication layers — and, for the campaign-sourced texts, the campaign's own Claude-based verifier agents rather than the competition's jury; no base of either corpus was verified line by line by a human (an author read those written by GPT, Kimi and GLM models and found them consistent with published solutions), so every "sound rejected" count below presupposes those model verdicts. One second-corpus base carries two claims its planting amendment records as asserted rather than proved; its label as an original is the evaluators' verdict, and the registered sensitivity arm (§B) excludes it. Most are not ours: 22 of the 36 bases across both corpora come from a public solving campaign released by a third party (Das, 2026), and the rest from our own runs. We redistribute the texts we plant into, with their digests, for reproduction; that release states no licence at the commit we cite (Ethics Statement). None of our conclusions concerns the campaign's own results. They are drawn from a pool of 65 such texts by a seeded rule fixed before any was inspected, with caps on authoring model and GPT-family share (§B); they cover all six problems equally and span 8 authoring models against 3 for the calibration bases, of which 6 of 8 were GPT-written.

Three mutations, one hunk each. Every base receives exactly three single-hunk edits. *Fatal* (F) makes a statement false and leaves unchanged a later line that uses it. *Harmless* (H) falsifies a displayed value nothing later uses (usually the next line continues from the correct one; a few are side computations or examples). *Gap* (G) deletes the argument for a load-bearing step and keeps its claim. Truth is fixed by the class table registered in the protocol: originals accept under both conventions, F and G reject under both, and H rejects as written but accepts under a repair-tolerant reading. The line between F and H is what the certificate can check — whether a later line *cites* the falsified statement — and not whether a reader could repair it: on some F texts the retained text supplies the repair and a reader applying a broader repair convention than ours can pass them, on others the certificate records that it does not. The release's log disputes 6 fatal edits (on calibration, a one-line repair restores the planted statement; on the 3 held-out ones, the retained text supports the consumer's conclusion) and 2 calibration deletions of a standard identity's derivation (a held-out one is kept). With the held-out ones set aside (post hoc), the held-out pair's strict mark, flag and audit accept 6/29, 7/29 and 4/29 broken texts and still reject 13/32, 11/32 and 8/32 sound ones, and §4 bounds the fatal labels by counting every fatal edit sound. This gives 72 mutants and 96 texts, 24 of each class, 16 per problem.

Falsity is checked, class is recorded, and the price is a judgement whose owner we name. The standing objection to planted defects is that the author decides which edit is fatal. One thing is taken out of our hands. Every F and H carries an *executable check* that refutes the planted statement by finite computation — 48 of them, all passing, with the 24 deletions correctly carrying none — so what is established is machine-checked falsity rather than opinion. Every mutant carries a *dependency certificate*: for F the quoted consumer line whose conclusion no longer follows, for H the quoted continuation showing that nothing downstream reads the display, for G the retained claim and the deleted argument; a reader checks the class structurally. A certificate is a frozen operational label, not a proof: it is model-written, and the executable check certifies that the planted statement is false, not that its consumer fails, which is how the disputed fatal labels arose. What no certificate fixes is the price. The rubric every grader receives bands a small, straightforwardly repairable error at 5–6 and is silent on the proof's own repair, so whether a cited false statement the retained text could repair costs a point or costs the proof is a judgement, not ours to certify. So, per class: falsity is machine-checked, class is structural, and the price is a judgement — the class table's, which is ours and registered as an instrument setting (F and G reject under both conventions, H rejects as written and accepts repair-tolerant), and the grader's, which §4 reads off the response and §5 prices against the same grader's grade on the same base's original.

Calibration and confirmation. The 32 texts planted first are the calibration set: the decision rules were chosen with them in view and they cannot test anything. The 64 new texts are the confirmation set, and no rule, threshold or convention that any reported endpoint rests on was added or altered after the first call against them, the matched audit prompt of §6 excepted, which is why its held-out figures are exploratory; the post-hoc readings this paper also prints are labelled as such where they appear, and no endpoint rests on them. Table 1 gives each rule on the confirmation set with the calibration figure beside it; counts over the whole corpus pool both sets, as their denominators show.

3 What a Grading Response Carries

The instrument is one ordinary grading call per text per lab. The grader is told it is a competition coordinator marking a single write-up on the standard 0–7 scale, is given the problem statement, and is asked to judge the text on its own terms rather than against a published solution. It returns a JSON object with four things we use: an integer `grade`, a boolean `complete`, a free-text `main_gap`, and a list of errors each carrying a description and a verbatim `quote` from the text. Nothing else about the call changes between labs, and the instruction block is byte-identical to the one used on the calibration set.

Every rule we compare is read off that single response. *Pass mark* k accepts when the score is at least k. *Strict* accepts only a 7. *Completeness* accepts when the same response set `complete` true, a free field. A combined rule accepts only if both labs of the decision pair accept. The obligation-level *witness* rule of §6 is the one procedure that costs a second call.

Two conventions, stated because they change the answer. A false displayed value the argument does not use — usually one the next line silently corrects — is an error under an *as-written* reading and not under a *repair-tolerant* one. We fix both and report both. The distinction is exactly the harmless class of §2, whose truth value differs between the two; §5 shows what turns on it. Beside these two frozen conventions, §4 reports a post-hoc label sensitivity: the same stored responses with every fatal edit priced sound.

4 Measurement Across a Grader Panel

The sweep. Over the 48 broken texts, a pass mark of 5 accepts 45 and a mark of 6 accepts 41; a 7 accepts 8 while rejecting 22 of 48 sound ones, where the completeness field, nested between 6 and 7, accepts 9 and rejects 17. The sweep bounds what any threshold on this rubric can do (Figure 1, §A, per grader).

The response locates what the score forgives. The sweep is not a claim that model graders miss planted errors. Of the broken texts a pass mark of 5 accepts for the decision pair, at least one grader quotes inside the planted hunk on 44/45. The pair carries it unevenly, 23/45 for one grader and 43/45 for the other, and the median grade of 6 is taken over the pair's grades on the located texts. The one accepted broken text on which neither quote reaches the hunk is a disputed fatal edit of §2: both graders returned full marks with the flag set and no listed error, and their notes say the retained case hypothesis supplies the step. So the grader read the edit, wrote it down and, at the median, gave the proof the grade the rubric's band prescribes for a repairable error; what a pipeline would need to price the edit is in the response its rule discards. At the strict mark the error list splits by grader: of the held-out broken texts each awards 7, one lists no objection inside the hunk on 13/13 and the other on 2/11, though the first grader's free-text notes often name the omitted step and call it routine, a price written outside the error list.

The pricing, not the reading. Nor does the pass mark's permissiveness depend on how a cited slip is priced (post hoc; no response re-graded). With every fatal edit counted sound and only the held-out deletions broken, the pass mark, flag, strict mark and audit accept 15/16, 6/16, 5/16 and 2/16 deletions and reject 0/48, 25/48, 27/48 and 22/48 sound texts: the rules keep their Table 1 order on both axes.

The comparison is per lab, and it does not go one way. On the held-out set, taking each lab's own response, the completeness field dominates the strict mark for 1 of 5 graders, the strict mark dominates the field for 2, and for 2 neither dominates: one accepts fewer broken texts and the other rejects fewer sound ones. On the stored run a pooled pair would say that no pass mark reaches the field's operating point; that is false for the 2 labs where the strict mark is better on one error kind and no worse on the other, and we report the per-lab table rather than the pooled pair that hides it (Table 2); 1 of the 5 verdicts is taken over what its lab returned, Kimi K3's 58 of 64 texts.

A single call does not hold a per-lab ordering. Each verdict rests on one or two texts. Re-run unchanged for the pair's graders on 2026-09-21 (stored run 2026-09-08; Table 2), the grade prompt changes the strict decision on 15/64 and 8/64 held-out texts, the first grader both ways, the second only toward 7 (13 up, 0 down over the corpus): its strict mark moves from 11/32 to 19/32 broken accepted, and the first grader's verdict turns from the field to the strict mark. The protocol registered this retest as the grade arm's noise floor (about three flips per axis expected); applied here post hoc, per-lab differences below a lab's flip count on the same axis are ties. The audit and the other graders were not retested (one lab's partial re-route set the expected floor).

On held-out texts the gap to a permissive mark is wide, and the gap between the alternatives is not. Table 1 gives the operating points. Because every rule is read off the same texts, the comparison between two of them is paired, and an exact McNemar test on the discordant pairs is the instrument, not the overlap of the marginal intervals. Against the pass mark the completeness flag is decisive on broken texts (23 against 0 discordant, $p < 0.001$), and so is the obligation audit (27 against 0 discordant, $p < 0.001$); the pass mark's compensation is that it rejects no sound text at all, which the flag (0 against 11 discordant, $p < 0.001$) and the audit (0 against 8 discordant, $p = 0.01$) both give

Decision rule	broken accepted	sound rejected
pass mark ≥ 5 (calibration 14/16, 0/16)	31/32 (97%, 84–99)	0/32 (0%, 0–11)
strict mark $= 7$ (calibration 1/16, 9/16)	7/32 (22%, 11–39)	13/32 (41%, 26–58)
completeness flag, same call (calibration 1/16, 6/16)	8/32 (25%, 13–42)	11/32 (34%, 20–52)
obligation audit, no answer key (calibration 0/16, 6/16)	4/32 (13%, 5–28)	8/32 (25%, 13–42)
main_gap = none, same call (post hoc; calibration 1/16, 3/16)	6/32 (19%, 9–35)	10/32 (31%, 18–49)
empty error list, same call (post hoc; calibration 0/16, 11/16)	1/32 (3%, 1–16)	22/32 (69%, 51–82)

Table 1: Every decision rule on the 16-base held-out set, decision pair, repair-tolerant convention. Parenthesised: the proportion and its 95% Wilson interval. Broken texts are the fatal and gap mutants; sound texts are the originals together with the harmless-display mutants, by the registered labels of §2, not by verified proof status. The two columns are errors in opposite directions, so a rule is better than another only if it is lower in both. The obligation audit is lower in both than the strict mark and the completeness flag — for this pair, not for each lab in it, and not than the permissive pass mark, which rejects no sound text by accepting almost every broken one; the audit's held-out point is exploratory and unretested. The intervals are marginal and treat texts as independent; the paired tests are in the text. Row labels carry the same two counts on the 8-base calibration set; the two rows under the rule read a free field of the same response post hoc and are not in the frozen set.

up. Among the three non-permissive rules nothing is settled: the audit-versus-comparator tests, two comparators by two error axes, run from $p = 0.18$ (the strict mark on the sound axis) to $p = 0.45$ (the strict mark on the broken axis and the completeness flag on the sound axis), and the two comparators agree on 61/64 of the held-out texts, so this is one comparison repeated on each error axis rather than four. A two-sided sign-flip permutation over the 16 held-out bases, which treats the texts of a base as one unit, gives $p = 0.53$ and $p = 0.18$ for the audit against the strict mark (post hoc), and the order survives the disputed labels' removal (§2), which does not certify the labels that remain. We report a consistent direction and an underpowered sample, not a ranking.

The field differs from the strict mark only within grade 6. Over the decision pair's 184/192 grading responses the completeness flag agrees with "grade $= 7$", and per lab the agreement runs from 71/85 to 95/96. The disagreement is one-sided and confined to one grade. Over all 469 first-corpus gradings from the 5 graders, every response graded 7 sets the flag (284/284), every response graded 5 or below clears it (0/41), and only grade 6 splits, at 27/144; for the pair alone the same counts are 103/103, 11/11 and 8/78. The flag is a strict refinement of "grade ≥ 6" and a strict relaxation of "grade $= 7$", nested between two thresholds, so whatever it adds to the score lies inside grade 6; held out, that is 2/30 of one grader's grade-6 texts (harmless and original) and 2/17 of the other's (fatal and gap), and the prompt fixes the scale but never says what the field means. The free gap field, read as the exact sentinel `none`, has the same shape (282/284 at 7, 0/41 at 5 or below, 12/144 at 6), and an empty error list is a refinement of the top mark alone (193/284 at 7, 0/144 at 6, 0/41 below). Where the flag lands between those cuts is the finding: against the strict mark on the held-out set it trades on the stored run, 8/32 broken accepted against 7/32 and 11/32 sound rejected against 13/32. What no threshold can do is separate the classes the conventions are about. Among the 34 held-out texts whose lower pair grade is 6 sit 14 fatal, 7 deletion, 10 harmless and 3 original texts: a cut between 6 and 7 rejects all 34 of them and a cut at 6 accepts all 34, so on this set no threshold on the pair's score separates a repaired line from an unrepaired one, whichever convention is in force.

5 Convention Identifiability

We call a false displayed value that the next line silently corrects, and no later line cites, a *silently repaired error*; it is the core of the harmless class of §2, whose few unused asides share the property that matters here, that nothing later reads the display, and whose certificate records a fact a reader can check rather than a matter of taste. A false display a later line does cite is the fatal class, whatever a reader could do to repair it, so the repair-tolerant convention we compute tolerates uncited slips alone. Only that class turns on the choice, and it is enough to decide which lab looks better. In the stored run, counting each lab's disagreements with the intended truth, the ranking reverses between the two conventions on all 3 strict rules — the top score, the completeness field, and the obligation-level

witness rule, which agree pairwise on 76–90 of 96 texts. Under the repair-tolerant convention Claude Opus 5 is the most accurate lab on every one of them; under the as-written convention GPT-5.6-Sol is, on every one of them. The same responses, the same texts, opposite conclusions. On the retest (§4; post hoc) the harmless texts at 7 stay apart (24/24 against 5/24; stored 23/24, 6/24) and the strict mark's reversal holds (errors 25/96 against 33/96 repair-tolerant, 49/96 against 19/96 as written); the flag's does not (37/96 against 30/96), and the audit was not re-run. The cleanest form of the difference is on the same texts: of the 15 harmless texts on which both graders quote the planted slip, GPT-5.6-Sol still awards 7 on 2/15 and Claude Opus 5 on 14/15 (12/15 split 6 against 7); the objection is the same, the price is not (pooled over both strata). The unedited base is the control (post hoc): GPT-5.6-Sol grades the original 7 on 11/15 of those 15 texts and the harmless mutant one point below its own original on 10/15 (8/15 on both runs); Claude Opus 5 grades the original 7 on 15/15 and the mutant below it on 1/15; over all 24 harmless mutants, 15/24 against 1/24. On 9/13 of the texts where the stricter grader withholds the 7 the slip is the only error it lists and its original stands at 7: the price is the slip's, not the proof's, and the comparison needs the certificate and the executable check, and no verdict on the correctness of the rest of the proof beyond the certificate's dependency claim (§J). Within each of the six problems this same-text difference runs the same way on the stored run ($n = 4, 3, 2, 2, 1, 3$ per problem; a one-text margin on P3 and P5), and a two-sided sign-flip over the six problem clusters gives $p = 0.03$, the smallest value six clusters can give under that test (§K). In §J's text a sign slip in a display draws a 6 from one grader ("the correction is immediate") and a 7 from the other ("harmless"). We read this as a repair convention, a supported interpretation rather than an identified cause: the no-repair arm of §6 is a partial instruction control against the silent plain prompt, and no prompt permitting the repair was run.

The choice is not academic, because the graders have already made it, each differently. On the harmless class, a grader that quotes the planted slip awards 7 anyway on 3/20 of the texts where it does so for GPT-5.6-Sol, and on 17/18, 11/11 and 11/14 for Claude Opus 5, GLM-5.3 and Kimi K3; the fifth lab quoted the slip on too few texts to rate, 1/3. Under the matched audit on the first corpus, whose prompt says not to repair, one grader accepts 24/24 harmless mutants and the other 12/24. Whether a proof with a silently repaired error is a proof belongs to the discipline, not to a grader; what the paper can say is what each answer buys. Count such errors sound, and no threshold on the score serves while the grader that repairs looks more accurate; count them wrong, and the maximum mark serves while the grader that reads as written does. A pipeline that has not scored for the choice has let its grader make it, and the graders make it differently under one prompt; this is the part of validation that does not delegate.

The consequence for practice is narrow: a reader of a verifier comparison that does not state its repair convention cannot tell whether a reported advantage is a property of the grader or of an unstated decision about what counts as an error. We report both throughout.

6 Routing at a Fixed Inspection Budget

Read each rule as a router: accept when both labs agree to accept, reject when both agree to reject, and refer the rest to a human. The audit is run as a comparator, not as a proposal: a second call with a stated no-repair convention, so the free rules have a point the score alone cannot produce to be read against; as a router it refers 24/96 texts, admits 4/48 broken ones and turns away 0/48 sound ones, against the field's 31/96, 9/48 and the same none. As a binary gate on the held-out set the ordering is the same, 4/32 against 8/32 broken accepted and 8/32 against 11/32 sound rejected. It is the one rule here that is not free; on natural proofs, model disagreement and human uncertainty pick out different texts (§H).

Every witness figure this paper reports as a finding comes from an information-matched arm whose prompt carries no required-mechanism list, no stated answer and no named false lemma: by content-word overlap against each planting, discounting the problem statement every arm receives, it discloses the planting on 0/72 mutants and the ordinary grade prompt on 0/72 (the measure covers 40/48 of the broken texts, since word overlap cannot see a short symbolic edit). It was built after the held-out calls, because the arm we ran first carried the problem's marking scheme (22/72 mutants disclosed; §E), so §2's freeze forbids presenting its held-out figures as confirmed, which is why §7 exists.

The prompts match on information, not on wording. The audit prompt is 1.8–2.4× longer and says not to repair the argument for the author, where the grade rubric puts a straightforwardly repairable error at 5–6; conversely the grade prompt bans consulting a solution or a lemma list (6/6 problems) and the audit prompt does not (0/6). The operating points bear on the reading that the audit is only a stricter grade prompt: a strictness shift moves a rule along the scale's own frontier, and the audit's held-out point is off it, 4/32 broken accepted against 8/32 sound rejected where the strict mark is 7/32 against 13/32. In these counts the audit lies below the strictest available cut on *both* axes, where no relabelling of a cut can place a rule, on discordant pairs too few to settle and with no retest of its own. With the audit's discipline sentences in the grade prompt and nothing else changed, the pair's strict mark accepts 6/32 held-out broken texts and rejects 13/32 sound ones, against the audit's 4/32 and 8/32 and a retest of the plain prompt run alongside the arm at 7/32 and 12/32. The instruction moves the strict mark along the scale's own frontier — 3 broken texts leave 7 (2 against the retest, inside its flips) while 13/32 sound ones are rejected — and the audit's point stays off it: no instructed pass mark, flag or gap field accepts as few as 4/32 broken texts while rejecting as few as 8/32 sound ones. Per lab: GPT 7/32, 13/32 (gap class 6/16 against the audit's 2/16); Claude 17/32, 0/32. Exploratory on the held-out split (the arm was registered after the confirmation calls); calibration 1/16, 9/16. Counting the reduction obligation the audit elicits but does not score (post hoc) moves its point from 4/32 and 8/32 to 3/32 and 12/32: against the completeness flag of Table 1, 8/32 and 11/32, fewer broken texts accepted and more sound ones rejected.

7 A Pre-Registered Replication, and What It Says the Rule Is

Because the matched arm was built after the held-out calls (§6), we built a second corpus and fixed its reading before it existed. 12 further proofs were selected from the same pool by a seeded rule, disjoint from the 24 already used and absent from the manifest, and given the same three edits under the same specification, for 48 texts. Two arms ran on them, the pass mark and the matched witness rule, on the same evaluator pair; the threshold was registered as a count before any text was selected: at most 6 broken texts accepted would confirm, more than twice that would fail.

It confirms the registered gap to the pass mark. On the second corpus, which the instrument had never seen, the matched arm accepts 4/24 broken texts and rejects 7/24 sound ones, where the pass mark accepts 22/24 and rejects 0/24. The gap this confirms is the one against a permissive pass mark; the threshold was not set to discriminate the audit from the free rules of §4, and the replication's pre-registration (§B) forbids reading those off this corpus as a confirmation: the strict mark accepts 6/24 broken texts here, which clears the registered bar too, so a reader should not take this replication as evidence that the audit beats the free rule. As a disclosed post-hoc read and not a second endpoint, the two-axis ordering matches the first corpus — the mark rejects 12/24 sound texts against the audit's 7/24.

What replicates is the grader, not the rule. The rule's operating point is not a property of the rule. Table 2 splits the audit by decider on the held-out set, and the second corpus repeats the split: one grader accepts 4/24 broken texts and rejects 7/24 sound ones, the other 12/24 and 0/24, rejecting no sound text on either corpus. The same disposition shows on the class the conventions of §5 are about: under the audit's own prompt, which says not to repair, one grader accepts 24/24 of the first corpus's harmless-display mutants and the other 12/24 (under the audit's accept rule, over all 24; an ordering, not a rate comparable to §5's). A grader that passes a display it can repair is applying a repair convention the prompt did not grant it. This is an observation, not a measured stability: a per-lab reading the replication's pre-registration excludes from the endpoint, on a second corpus half the size of the first, with the second grader's acceptance rate moving by 12 points between them. What two corpora establish is that the gap between the graders does not close; and since the combined rule accepts only when both accept, the pair inherits the stricter grader's operating point.

This is the same dependence the panel showed for the completeness flag. For the decision pair, an obligation-level rule is better on both axes than the strict mark and the completeness flag — not than a permissive pass mark, which rejects no sound text and buys that by accepting almost every broken one, and, as §F shows, not evenly across the mutation classes. But that ordering is itself a property of the pair, and split by lab it does not hold still: in counts, for one grader the audit is below either alternative on both axes, while for the other the strict mark accepts 11/32 broken texts on the stored

	GPT-5.6-Sol		Claude Opus 5		GLM-5.3		Kimi K3		DeepSeek-V4-Pro		**pair**	
Decision rule	acc.	rej.	acc.	rej.	acc.	rej.	acc.	rej.	acc.	rej.	acc.	rej.
pass mark ≥ 5	31/32	0/32	31/32	0/32	30/32	0/32	27/29	0/29	23/32	3/32	31/32	0/32
strict mark $= 7$	13/32	13/32	11/32	0/32	14/32	0/32	11/29	1/29	13/32	6/32	7/32	13/32
completeness flag	13/32	11/32	13/32	0/32	15/32	0/32	21/29	0/29	15/32	4/32	8/32	11/32
obligation audit	5/32	8/32	12/32	0/32	–	–	–	–	–	–	4/32	8/32
the same grade prompt re-run on 2026-09-21 (stored run: 2026-09-08)												
strict mark, retest	9/32	12/32	19/32	0/32	–	–	–	–	–	–	7/32	12/32
flag, retest	10/32	12/32	26/32	0/32	–	–	–	–	–	–	9/32	12/32
post hoc, not in the frozen set												
main_gap $=$ none	13/32	10/32	10/32	0/32	13/32	0/32	12/29	1/29	14/32	5/32	6/32	10/32
calibration set	4/16	3/16	2/16	1/16	6/16	0/16	2/13	2/14	5/16	0/16	1/16	3/16
empty error list	13/32	16/32	2/32	21/32	7/32	13/32	6/29	11/29	12/32	8/32	1/32	22/32
calibration set	3/16	9/16	0/16	8/16	6/16	3/16	2/13	8/14	5/16	1/16	0/16	11/16

Table 2: Every rule read off one grading response, per grader and for the decision pair (both must accept), on the 16-base held-out set, repair-tolerant: broken texts accepted (fatal and gap mutants) and sound texts rejected (originals and harmless-display mutants), over the texts each decider returned (Kimi K3: 58 of 64). The audit ran on the pair only. The retest rows re-run the unchanged grade prompt; a per-grader difference below that grader's strict flips on the same axis (broken, sound: 8/32, 7/32; 8/32, 0/32) is read as a tie, the grade arm's registered floor applied post hoc. The rows under the last rule are post hoc, their calibration counts beneath them.

run and 19/32 on the retest to the audit's 12/32, all rejecting 0/32 sound ones. Outside a grader's retest flips lie only the first grader's stored broken-axis margins and the second's retest margin against the flag (26/32 to the audit's 12/32); none holds on both runs, so per lab we report no stable ordering. And the ordering turns on the convention as well as the grader: counted as written, on the held-out set, the audit accepts 14/48 broken texts and rejects 2/16 originals against the strict mark's 13/48 and 3/16 — a trade, not a dominance. Which decision rule is best is not a fact about the rules alone.

8 External Validation on Human Labels

Every number so far rests on labels we created. This section repeats the comparison against labels we did not: the Open Proof Corpus (Dekoninck et al., 2026), 644 human-judged proofs of 306 problems from 17 competition series as the corpus labels them (1977–2024), drawn from its generic split, 616 of them carrying a majority verdict (225 correct, 391 incorrect) and 28 split evenly with no majority at all, which we exclude from every accuracy figure rather than break toward either side. Two strata were fixed before any call: a targeted one holding every multi-judged and every span-annotated proof, and a random one of 200 proofs. Every error rate of a decision rule is taken from the random stratum. Readings that need more than one judge or a human error span (the majority counts below, §H) come from the targeted stratum and are properties of the human labels, not rates of ours.

On natural proofs the flag again differs from the thresholds only within grade 6. On the random stratum, among rules that accept no incorrect proof, the strict mark costs 39 false rejects of 200 and the completeness flag — nested between the same two thresholds as on the planted suite, at a point no threshold reaches, which the protocol's registered reading counts as a transfer — costs 35, its accepts strictly containing the mark's; under equal weights a pass mark of 5 errs on 20 proofs against the flag's 35. The honest statement is conditional on the operator's loss, and on this sample: if no incorrect proof may be accepted the flag errs least here; if the two errors weigh alike, the pass mark errs less.

Discrimination does not transfer. On these proofs the grader whose planted-corpus area is lowest (GPT-5.6-Sol) separates human-incorrect from human-correct at 0.95 against 0.97 for the other, and mutation testing measures what an observational corpus cannot (§M).

The rules against the human majority, as counts on the rows that have one. The protocol pre-registers every rule's error on the multi-judged proofs as an endpoint, reported as counts, not a rate, because every multi-judged proof is in the targeted stratum ("No rate computed on Stratum A is a population rate"). Of the 282 proofs judged by more than one person, 254 carry a majority label (254/254 of them in the targeted stratum), over 59 problems, and on those the pass mark errs on 11 (6 accepted against the majority, 5 rejected against it), the strict mark on 52 (1 and 51) and the flag on 47 (1 and 46) — almost only by rejection, and partly by definition, since the corpus's judges were instructed to call a solution correct "even if it would earn 5+/7 points" with "a small computational mistake that can be easily fixed" worth a one-point penalty (Dekoninck et al., 2026), so the label is itself a pass mark of 5 by instruction. The split by unanimous and divided rows is in §M. Any claim that a rule is *wrong* on this corpus has to reckon with a target on which the judges themselves differ on 33 of 282 rows.

Limits. The corpus has been public since June 2025 and these graders are 2026 models, so their judgments may be in training data; that discounts raw accuracy, not a grader's own score against its own field. The discrimination AUC and the uncertainty-flag count pool both strata (§M, §H).

9 Related Work and Limitations

Reference-guided grading of competition proofs, validated against human marks, is established (Mahdavi et al., 2025; Ma et al., 2025); so is that an LLM judge is moved by candidate order, verbosity, its own authorship and trivial tokens (Wang et al., 2024; Zheng et al., 2023; Panickssery et al., 2024; Zhao et al., 2025) — biases in how a judge scores, where what we report is prior to scoring: an unstated standard of what counts as an error, which two graders resolve differently under one prompt (self-preference, the one such bias our design can test, is §G). Petrov et al. (2025b) plant falsity into the *problem* and ask whether a prover objects; we perturb an *accepted proof*, in a way an executable refutation settles, and ask what the grader's own output establishes about it. Step-level error labels (Lightman et al., 2024) and first-error benchmarks (Zheng et al., 2025) ask where an error lies; we ask what a located error costs. Planted errors test critics and judges already (McAleese et al., 2024; Doddapaneni et al., 2024), located natural ones too (Tyen et al., 2024), and proof autograders are validated against human grades (Luong et al., 2025); we add an executable refutation of each planted falsity and ask what the grade prices. The external corpus of §8 builds the permissive pricing into its labels (a pass mark of 5 by its judge instructions; Dekoninck et al., 2026, App. H). That models state confident proofs of what they have not proved is solver-side and established separately (Petrov et al., 2025a), as are low proof scores on contest problems (Balunović et al., 2025); §N records our own instance of the former. Joint generation, verification and repair (Chen et al., 2026), verification-and-refinement at contest scale (Huang & Yang, 2025) and expert grading of research-level proofs (Schmitt et al., 2025) are the complementary directions; the Open Proof Corpus (Dekoninck et al., 2026) supplies the human labels of §8. What we add is not a better grader: a corpus in which the planted falsity is machine-checkable, the finding that the standard of correctness a grader applies under one rubric is its own and that the audit prompt's stated convention bound one grader of the pair only partly and the other not at all, and decision rules read against the rate at which human judges disagree with each other.

What we do not claim. *Whole-proof correctness:* 72 executable checks certify the planted statements false, the certificates record a dependency, and the bases carry model verdicts; §5's same-text comparison needs no verdict on the correctness of the rest of the proof, its control being the same grader's grade on the same base's unedited original and its class the certificate's dependency claim; Table 1's sound axis rests on the bases' model verdicts: 4/16 of the 16 held-out originals are rejected by the strict mark, the flag or the audit (§L), and a rejected harmless mutant is an error only if its base is sound. *Breadth:* six IMO 2026 problems, 36 bases; the direction of §5 holds within each of the six on the stored run (§K), no planted result generalises beyond one olympiad, and §8 tests the flag's nesting and the rules' errors against a human majority, not the pricing, on 306 natural problems (59 in the multi-judged, majority-labelled subset; counts, never a rate). *A superior grading method:* the audit is a comparator, run on two graders, unretested, its gain in the announced deletions (§F, post hoc), its per-lab order unstable (§7). Not a stable per-grader operating point: one retest moved it by more than most differences we report (§4). Not independent texts: a base's texts share most of their content, which the paired tests ignore; and no human coordinator priced the mutants.

REPRODUCIBILITY STATEMENT

Every count, rate and interval over the corpora that the paper writes as a figure is a LaTeX macro that an analysis script writes from the stored responses, so such a figure cannot disagree with the computation behind it. Apart from the words "one" and "two", the literals the prose does carry are scale values and thresholds, the confidence level, dates, model versions, design constants such as the three mutation classes, prose fractions, the grade-arm protocol's expected retest floor, and one incident count from the deviations log (§C); each is registered, with its reason, in `scripts/literal_check.py`, which fails on an unregistered one. That guard admits a short list of small numerals and number words in any context and does not scan math mode or the argument of a one-argument command, so it does not check the small counts written those ways, such as the accepted broken text of §4 that no quote reaches. The released bundle carries every grading response from both corpora, all 144 texts with their digests, the planting specifications with their diffs, the executable refutations, and the analysis scripts (Python, with sympy for the refutations and matplotlib for the figure); re-running them offline regenerates every macro and the figure except those computed from the external corpus's proofs (§8), whose corpus we may not redistribute — for those we ship our per-proof decisions, the identifiers and digests needed to rebuild the sample, and the code, but not the proofs themselves. The solver-side note of §N ships the per-run record of every run it counts — configuration, first message, final status file and snapshots, the verifiers' reports and code, the tool-call audit — with the timestamped protocol amendments that governed them and the script that derives its figures from them; the full trajectories are withheld from the bundle for size. An independent checker re-derives each corpus property from the files rather than trusting the manifest: the selected bases' digests, single-occurrence search strings, edits confined to their span, class-consistent truth values, and a passing refutation for every fatal and harmless mutant. Protocols for base selection, planting, the held-out split, the external corpus, the replication (its pre-registration and amendment) and the discipline-matched grade arm were written before the calls they govern (most are dated to the day only, §B), and every departure is recorded with its date, most also with a UTC time, in the released deviations log, including the ones that were mistakes. The procedure itself — decide the convention in writing before any call, build a probe set whose falsity is machine-checked, run the rules under both conventions and report both error axes per grader, then check whether the grader honoured the convention it was told — is released in the bundle as a skill file an agent can load and a checklist a reader can follow, with the failure modes we hit recorded beside the steps that produced them.

ETHICS STATEMENT

The study involves no human subjects and no personal data. 22 of the 36 base proofs come from a public solving campaign whose release states no licence at the commit we cite (Das, 2026); we redistribute those texts and their mutants only so that this study can be reproduced, name the source, and ask anyone who intends to reuse them to seek terms from it (`NOTICE.md`). The proofs of the external corpus of §8 are not redistributed beyond the excerpts graders quoted in stored responses; those excerpts and our derived decisions carry its licence.

USE OF AI ASSISTANCE

Language models are the object of study here, and they were also used extensively to produce the work. Against the categories in the ICLR 2027 AI Policy for Authors: models *implemented methods* (the analysis, plotting and transport code, and so the figure, were largely model-written under author direction), *designed and gave feedback on methodology* (protocol drafts, including the held-out split and the delivery diagnostics), *assisted in the writing* (this manuscript was drafted by a model and revised by the authors), *cleaned and reformatted data*, *helped interpret results*, and *proposed and refined hypotheses*: the split of the audit's gain by mutation class and the announced-deletion reading of §F came from a model reading the stored replies (recorded in the deviations log), while the initial idea, the study design and every experiment were the authors', who directed the models' execution and analysis and validated the results stage by stage. Models also wrote the dependency certificates and the executable refutations the labels rest on, which is why §2 names what each establishes and the release names the certificates found wrong. The conventions, the class table and the endpoints were fixed in protocols the authors approved before the calls they govern. Models were not used for

translation, and the paper states no theorem of its own. The planted mutants were authored by models working from a written specification, one per group of base proofs; their output was not accepted on report but re-verified independently by the checks described in §2, and the executable refutations are what make the planted falsity trustworthy; the class labels rest on the certificates, and the release names the ones found wrong. In §N the adjudication of claimed proofs was itself carried out by model verifiers writing executable checks, and the section says so rather than presenting it as human verification. No model is an author.

The authors are responsible for the contents. Two safeguards exist because of that responsibility rather than in place of it: every count, rate and interval over the corpora is generated from the released artifacts rather than transcribed, with the exceptions the Reproducibility Statement states, and an automated gate checks that each guarded figure names the population it is over and appears beside its paired outcome. Neither substitutes for author review, and several errors reported in the deviations log were caught only by re-reading the data.

REFERENCES

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.

Jiacheng Chen, Xinyu Zhang, Shunkai Zhang, Yanmohan Wang, Lin Li, et al. MaxProof: Scaling mathematical proof with generative-verifier RL and population-level test-time scaling. arXiv:2606.13473, 2026. Submitted June 11, 2026.

Debarghya Das. imo-2026: Kimi K3 on the 2026 International Mathematical Olympiad. GitHub repository, commit `dbe8723`, `https://github.com/deedy/imo-2026/tree/dbe872307883296c5f06bf8bcb90e6bfd0879364`, 2026. Public campaign; scores from Claude-based verifier agents, not the IMO jury.

Jasper Dekoninck, Ivo Petrov, Kristian Minchev, Mislav Balunović, Martin Vechev, Miroslav Marinov, Maria Drencheva, Lyuba Konova, Milen Shumanov, Kaloyan Tsvetkov, Nikolay Drenchev, Lazar Todorov, Kalina Nikolova, Nikolay Georgiev, Vanesa Kalinkova, and Margulan Ismoldayev. The open proof corpus: A large-scale study of LLM-generated mathematical proofs. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.21621v2. Dataset: `https://huggingface.co/datasets/INSAIT-Institute/OPC`, CC BY-NC-SA 4.0.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. Finding blind spots in evaluator LLMs with interpretable checklists. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16279–16309, 2024.

Yichen Huang and Lin F. Yang. Winning gold at IMO 2025 with a model-agnostic verification-and-refinement pipeline. *arXiv preprint arXiv:2507.15855*, 2025.

IMO. Problems of the 67th International Mathematical Olympiad. `https://www.imo-official.org/problems/`, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Thang Luong et al. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Wenjie Ma, Andrei Cojocaru, Neel Kolhe, Bradley Louie, Robin Said Sharif, Haihan Zhang, Vincent Zhuang, Matei Zaharia, and Sewon Min. Reliable fine-grained evaluation of natural language math proofs. *arXiv preprint arXiv:2510.13888*, 2025.

Hamed Mahdavi, Pouria Mahdavinia, Samira Malek, Pegah Mohammadipour, Alireza Hashemi, Majid Daliri, Alireza Farhadi, Amir Khasahmadi, Niloofar Mireshghallah, and Vasant Honavar. RefGrader: Automated grading of mathematical competition proofs using agentic workflows. *arXiv preprint arXiv:2510.09021*, 2025.

Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz, and Jan Leike. LLM critics help catch LLM bugs. arXiv:2407.00215, 2024.

Moonshot AI. Kimi K3: Open frontier intelligence. Technical report, https://github.com/MoonshotAI/Kimi-K3/blob/main/k3_tech_report.pdf, 2026.

OpenAI. GPT-5.6 Sol model documentation. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026. Knowledge cutoff February 16, 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37*, 2024.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. Proof or bluff? Evaluating LLMs on 2025 USA math olympiad. *arXiv preprint arXiv:2503.21934*, 2025a.

Ivo Petrov, Jasper Dekoninck, and Martin Vechev. BrokenMath: A benchmark for sycophancy in theorem proving with LLMs. *arXiv preprint arXiv:2510.04721*, 2025b.

Johannes Schmitt, Gergely Bérczi, Jasper Dekoninck, Jeremy Feusi, Tim Gehrunger, et al. IMProofBench: Benchmarking AI on research-level mathematical proof generation. *arXiv preprint arXiv:2509.26076*, 2025.

Gladys Tyen, Hassan Mansoor, Victor Cărbune, Peter Chen, and Tony Mak. LLMs cannot find reasoning errors, but can correct them given the error location. arXiv:2311.08516, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2305.17926.

Z.ai. GLM-5.3 model card. Hugging Face, https://huggingface.co/zai-org/GLM-5.3, 2026. Repository created August 25, 2026; same base model as GLM-5.2; no knowledge cutoff published.

Yulai Zhao, Haolin Liu, Dian Yu, Sunyuan Kung, Meijia Chen, Haitao Mi, and Dong Yu. One token to fool LLM-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

A Operating Points per Grader

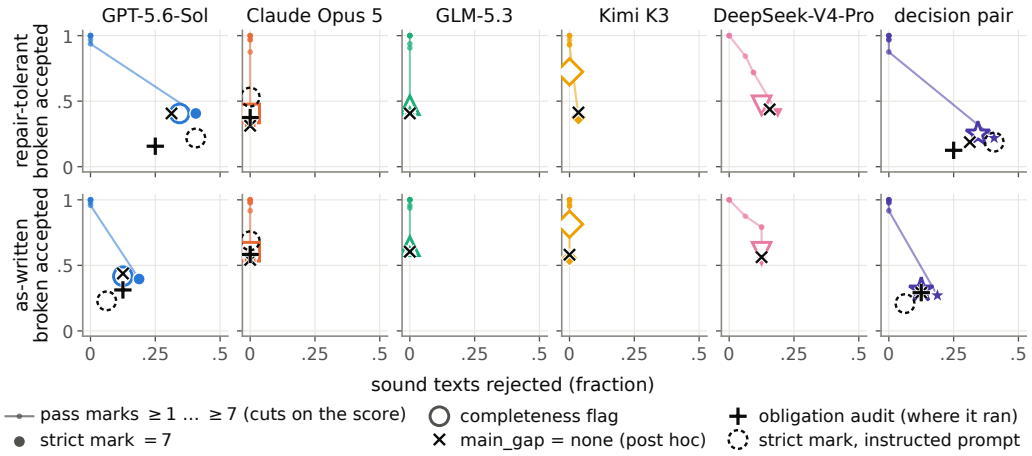

Figure 1: Operating points of every decision rule on the 16-base held-out set, one plot per grader and one for the decision pair (both must accept): repair-tolerant (top; 32 broken texts, 32 sound) and as written (bottom; harmless displays counted broken: 48, 16), as fractions of what each decider returned (Kimi K3: 58 of 64). The line joins the pass marks ≥ 1 to ≥ 7; the strict mark is its lower end. Touching markers are within two texts of each other (as written, the pair's main_gap and audit coincide: same counts, different texts); Table 2 has the counts. Dashed ring: the strict mark under the instructed grade prompt (exploratory).

B Protocols, Written Before the Calls They Govern

7 protocols are released with the corpus; the calibration planting protocol carries its launch time and the others the date on which each was written, most the date alone, so their order against the first call they govern on that day rests on the log and the stored request timestamps, not on the file. The grade-arm protocol records the minute it was signed and a digest of its text taken before its launch fields were filled; the release anonymises those fields and its signature line, so that digest cannot be recomputed from the released copy. **Base selection** fixes the eligibility rule, the seeded ordering, the per-problem quota and the diversity caps; its amendments are recorded in the same file, including the two allocator versions that starved a problem before the caps were satisfiable. **Planting** fixes the three mutation classes, their truth values under both conventions, and the requirement that every fatal and harmless edit carry an executable refutation. **Confirmation** fixes the calibration/confirmation split and the rules, combination, conventions and endpoints that may not change after the first call against the held-out texts. **External validation** fixes the sample, the strata, the instrument and the six endpoints for the human-labelled corpus, and states the licence and contamination limits before any result was seen. **The replication** of §7 has its own pre-registration and one amendment: the first fixes the pool, the seed, the caps, the two arms, the decision pair and the three thresholds before any base was selected; the second records the corpus properties the mutant authors found, the sensitivity arm excluding the two bases with a disclosed ease-bias, and the pre-call leakage check, all before the first call. On the second corpus, two of the 12 bases carry a disclosed property that makes a broken text easier to catch, a bias toward confirming, so the reading excluding them was registered too: over the remaining 10 bases the matched arm accepts 4/20 against 7/20 sound rejected, a rate of 20% against the 25% the registered count expresses, toward the threshold without reaching it and not a second verdict. Before any call we also checked, from the prompt sources rather than from stored requests, that neither arm in use shares content words with any planting on this corpus: the matched arm discloses 0/18, the grade prompt 0/18, while the keyed arm we did not run discloses 8/18. Word overlap cannot see a short symbolic edit, so that check covers 18/24 of the broken texts; the harmless mutants it cannot measure are sound under the convention in force and are not in the endpoint population. **The discipline-matched grade arm** of §6 has its own protocol, signed before its first call and before the retest of the plain prompt run alongside it, which it also governs, but after every stored call the arm is compared with, which is why its held-out figures are exploratory.

C Deviations

The released log records every departure from those protocols with its date, most also with a UTC time, including the ones that were mistakes: a transport swapped on a misdiagnosis and swapped back; an empty-reply retry that sent an option one model refuses, costing 22 texts; a retry condition that treated a truncated reply as a success; a lab that could not be served for its longest texts; and a sixth grader withdrawn after a partial delivery and never analysed. Two further mistakes, per-lab statistics that reported zero errors for a lab with no data and an analysis that read only two labs while 5 had run, are recorded in the comments of the analysis code where they were fixed rather than in the log. Each is recorded because a reader assessing this corpus needs to know how it was produced, not only what it contains.

D Per-Lab Delivery

The information-matched witness arm — the one every witness figure in this paper is taken from — was run on the decision pair only, 2 of the 5 labs, at 96 texts each. Every panel-wide figure in §4 and §5 is therefore a grade-gate figure, and this paper makes no panel-wide claim about the audit. On the grade and keyed-witness gates, 4 labs deliver all 96 texts. One delivers 85 of 96 on the grade gate and none on the keyed witness gate; its missing texts are longer than its delivered ones, so every figure involving it names the population it is over, and the panel is additionally reported on the 85 texts every lab completed.

E The Keyed Arm, Reported Only to Price the Answer Key

The arm we ran first was prompted with the problem's marking scheme, which on one problem names the planted false lemma outright. On the held-out set it admits 1/32 broken texts against 9/32 sound ones, where the matched arm admits 4/32 against 8/32. That comparison is not like-for-like and is reported only to show what the answer key was worth: 3 texts. Building the matched arm after the held-out calls was a change to the instrument after the confirmation set had been seen, which §2's freeze forbids presenting as confirmed; the second corpus of §7 is the confirmation.

F What the Audit Catches Inside the Gap Class

A post-hoc split of the held-out set, computed after the confirmation and reported as post-hoc, locates the audit's gain over the strict mark entirely in one mutation class. On fatal mutants the strict mark accepts 2/16 and the audit 2/16; on gap mutants, 5/16 against 2/16. For the grader that drives the pair the contrast is starker, 3/16 against 3/16 on fatal mutants — no gain at all — and 10/16 against 2/16 on gap mutants. The class where the audit wins is the deletion class, the one class carrying no executable refutation: its label rests on the dependency certificate alone. Inside that class the audit tracks something narrower than we would like. 13/16 of the held-out gap mutants replace the deleted argument with a sentence announcing the omission — that a step is standard, or that details are left out — because §2's specification fixes what a gap mutant must retain and says nothing about how the replacement reads, so the wording was a free parameter of the authoring model. The audit rejects 13/13 of the announced deletions and 1/3 of the 3 unannounced ones. With 3 texts we cannot say what it does on a deletion that keeps quiet about itself, and we do not claim to know. The defensible reading is that an obligation-level audit reliably catches a load-bearing step whose absence the text admits — a useful thing to catch, and a smaller thing than catching absence as such.

G Self-Preference on GPT-Authored Bases

On the held-out set, grading broken texts, GPT-5.6-Sol's strict mark accepts 5/8 of those from GPT-authored bases against 8/24 from the rest, and its completeness flag 5/8 against 8/24. The obligation audit shows no such gap, 1/8 against 4/24. With 8 texts on the own-author side this is not a finding about self-preference, and we do not report it as one; we report it because it runs in the direction that would flatter the two free rules, and because the audit, which the paper reads as a comparator, does not show it.

H Where Objections Land on Human-Marked Spans

Objections land where humans mark errors about a third of the time. The protocol pre-registers this on 222 annotated proofs; 15 of them carry no usable span, 17 annotations in all, so it is computed on the 207 that remain. On those proofs with usable sentence-level human annotations, 131/391 of one grader's located objections overlap a human-marked span (34%) and 172/550 of the other's (31%). The second quotes far more accurately yet lands on the human's span no more often, so quoting fidelity and diagnostic aim are separate abilities. 26 proofs carry a human uncertainty flag; only 2 of them are among those the two graders disagree about.

I The Gap Mutants That Show in the Numbering

7 of the 24 gap mutants, on 4 of the six problems, delete argument text that carries numbered equation tags, so the surviving numbering skips. A grader could in principle notice the deletion from the numbering rather than from the mathematics. The set is derived from the released texts by diffing each base against its gap mutant, not listed by hand. Removing a tag is not the same as deleting text that happens to contain one — two gap mutants carry a tag into their replacement and leave the numbering intact, and counting those inflates the figure. Every gap-class figure is reported with and without this set; on the 17 gap mutants that remain, the gap-class results are those of `policy.gap_without_tag_trace` in the first corpus's released `artifacts/planted_defects/results.json`.

J A Worked Example: One Harmless Edit, Two Prices

Every field below is copied from the release by `scripts/worked_example.py`: the planting, the certificate's quoted continuation, and each grader's stored response.

Text `p4-deedy-gpt-5.6-sol-pro-14f66095-H` (held out, problem 4). **Planted edit:** the display $N - b - x = N - k$ becomes $N - b + x = N - k$. **Retained continuation** (quoted by the certificate): "Moreover, $0 < k < N$, since $b > 0$ and $k < a + b = N - c < N$. Therefore k and $N - k$ are positive integers. Whichever triangle Shan-Yu retains, it has a positive integral angle, from which Mulan can force angle 1 by the induction above." — the value carried forward is the correct one, and no later line cites the planted display.

GPT-5.6-Sol: grade 6, `complete` false, `main_gap` "A sign typo makes one displayed angle calculation false, though the required correction is immediate from formula (1)." Listed error on the hunk: "The other child's integral angle is N-b-x=N-k, not N-b+x. With this sign corrected, both possible retained triangles have positive integral angles and the argument works." Notes: "Apart from this localized sign error, both sufficiency and necessity are rigorous and establish the stated classification."

Claude Opus 5: grade 7, `complete` true, `main_gap` "none" Listed error on the hunk: "Typo in the sufficiency step: the second child's angle is s-b-x = N-b-x = N-k, but it is written as 'N-b+x'. The asserted value N-k is nonetheless correct and follows immediately from formula (1) with x=k-b, so this is a harmless slip." Notes: "Correct answer 180°/N (N≥2 integer). Both directions are complete: integral-angle induction plus a clean existence argument for an integer in (b, a+b), and a valid four-case parity/integrality invariant for Shan-Yu starting from an equilateral triangle."

The base as the control, over the 15 same-text harmless mutants (post hoc). Grades on (original, mutant): GPT-5.6-Sol 10 at (7,6), 3 at (6,6), 1 at (6,7), 1 at (7,7); Claude Opus 5 14 at (7,7), 1 at (7,6). The one text graded above its original is `p1-deedy-grok-4.5-b031b8f1-H` (stored original 6, mutant 7; on the retest original 7, mutant 6). On the retest the mutant is graded below its original on 11/15 and 0/15 of the 15 (over all 24 harmless mutants, 18/24 and 0/24); on 8/15 GPT-5.6-Sol's drop holds on both runs (Claude Opus 5 0/15). On the 4 withheld texts where the slip is not the only error listed, GPT-5.6-Sol lists an error outside the hunk on 4/4 and Claude Opus 5 on 3/4.

K PER PROBLEM

Both tables are post hoc. The first splits the same-text harmless comparison of §5 by problem and by stratum, and adds the instructed-prompt arm of §6 on the same texts. The second gives Table 1 per problem (decision pair, repair-tolerant, held-out set). The two-sided sign-flip over the six problem clusters of the first table gives $p = 0.03$, the smallest value six clusters can give under that test.

	texts	GPT-5.6-Sol at 7	Claude Opus 5 at 7
P1	4	1	4
P2	3	1	3
P3	2	0	1
P4	2	0	2
P5	1	0	1
P6	3	0	3
calibration set	5	0	4
held-out set	10	2	10
all, instructed prompt (§6)	15	1	15

	pass mark ≥ 5		strict mark $= 7$		completeness flag		obligation audit	
problem	acc.	rej.	acc.	rej.	acc.	rej.	acc.	rej.
P1 (2 bases)	4/4	0/4	2/4	2/4	2/4	2/4	1/4	0/4
P2 (4 bases)	8/8	0/8	3/8	2/8	4/8	2/8	1/8	1/8
P3 (1 base)	2/2	0/2	1/2	0/2	1/2	0/2	1/2	0/2
P4 (4 bases)	8/8	0/8	1/8	3/8	1/8	2/8	0/8	2/8
P5 (3 bases)	6/6	0/6	0/6	4/6	0/6	3/6	0/6	4/6
P6 (2 bases)	3/4	0/4	0/4	2/4	0/4	2/4	1/4	1/4

L THE HELD-OUT ORIGINALS THE STRICT MARK, FLAG OR AUDIT REJECTS

Of the 16 held-out originals, 4/16 are rejected by the strict mark, the completeness flag or the audit of the decision pair; every field below is copied from the release by `scripts/worked_example.py`. Their labels as originals are the evaluators' verdicts (§2); the grader's objection is what the record holds, and no reason is inferred.

- `p1-deedy-grok-4.5-b031b8f1`: GPT-5.6-Sol: grade 6, `complete` false, `main_gap` "The claimed well-foundedness of the lexicographic order is false; termination needs the additional observation that P never increases when k decreases."; Claude Opus 5: grade 7, `complete` true, `main_gap` "none".
- `p4-deedy-muse-spark-1.1-220590cd`: GPT-5.6-Sol: grade 6, `complete` true, `main_gap` "none"; Claude Opus 5: grade 7, `complete` true, `main_gap` "none".
- `p5-deedy-deepseek-v4-pro-24b98fb5`: GPT-5.6-Sol: grade 6, `complete` false, `main_gap` "The stated case split omits the possibility that g takes only the positive value c; the subsequent contradiction argument must instead treat whether 0 is also attained."; Claude Opus 5: grade 7, `complete` true, `main_gap` "none"; GPT-5.6-Sol audit upper bound: written but invalid.
- `p5-ours-nemotron-28cf6b5e`: GPT-5.6-Sol: grade 7, `complete` true, `main_gap` "none"; Claude Opus 5: grade 7, `complete` true, `main_gap` "none"; GPT-5.6-Sol audit lower bound: written but invalid.

M DISCRIMINATION, AND THE SPLIT HUMAN MAJORITY

On the planted corpus. Across 5 labs from as many organisations — OpenAI (OpenAI, 2026), Anthropic, Z.ai (Z.ai, 2026), Moonshot (Moonshot AI, 2026) and DeepSeek — 4 of them complete on all 96 texts, the area under the curve separating a fatal edit from a cosmetic one runs 0.56, 0.71, 0.77, 0.80, 0.83, worst for GPT-5.6-Sol and best for Claude Opus 5; on the 85 texts every lab completed it

runs 0.56, 0.71, 0.79, 0.79, 0.80. For the 4 labs whose fatal and harmless grades all lie in $\{6, 7\}$ the area is exactly $\frac{1}{2} + [P(H{=}7) - P(F{=}7)]/2$, so the spread orders those graders by how much more often they award 7 to a harmless display than to a fatal edit, largely the convention of §5 rather than a discrimination; we withdraw it as a ranking.

On natural proofs. On planted mutants one grader's fatal-versus-harmless area is near one half because its fatal and harmless grades both lie in $\{6, 7\}$, so the area reduces to the gap between the 6/24 harmless and 3/24 fatal texts it grades 7 — a property of the convention it applies rather than of its discrimination. On these proofs the same grader separates human-incorrect from human-correct at 0.95, against 0.97 for the other. A planted fatal edit is a false line inside a sound proof; a naturally incorrect proof is usually wrong throughout, and a grader that tracks global quality scores both correctly. Mutation testing therefore measures something an observational corpus cannot, which is the case for §2 rather than an apology for it. On the random stratum alone the two areas are 0.94 and 0.98, against the pooled 0.95 and 0.97; the area needs every labelled proof there is, so the pooled figure is the one reported.

The human majority, split post hoc. On the 249 unanimous rows of §8 the pass mark, strict mark and flag err on 11, 50 and 45; on the 5 rows where the judges split two to one the pass mark sides with the majority on 5/5 and the strict mark and flag on 3/5 and 3/5; the 28 two-judge ties are outside the population, and on them every rule agrees with exactly one of the two judges.

N A Solver-Side Note

Some of our bases come from our own solving runs (§2). Before counting one of those runs' recorded failures — an open-weight model on one problem — we ran controls. The recorded run had ended in the agent client's output-cap error, not in a proof. Of the 13 control runs that declared completion in their own status file, 4 did not survive checking: 3 were refuted, two by model verifiers and one by its wrong final value, and one contradicted itself. A solver's "completed" is a field of its response, like a grader's completeness flag, and no more bound by the instruction that defines it. The controls compare no models and support no solving rate: the verifiers are language-model agents, a closure they accepted may rest on a step they repaired, and runs that claimed nothing were not checked. The per-run records, the verifiers' reports and code, and the timestamped amendments that governed the runs are in the bundle (`artifacts/p3_casestudy/`), so the reading can be repeated.