UNDERSTANDING DISAGREEMENTS IN LEAN FORMALIZATION EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating the faithfulness of a Lean formalization requires interpreting both its natural-language source and its formal statement. LLM judges, human reviewers, and published labels can disagree because of translation defects, mistaken judgments, or differences in interpretation and grading criteria. We investigate these disagreements using three LLM judges on CriticLeanBench, MA-Align, and 208 formalizations with existing human ratings. An author audit of 75 CriticLeanBench items includes all 51 cases where the judges unanimously contradict the label: the audit supports the judges in 44 cases, the label in 2, and leaves 5 unresolved. Source comparisons and targeted Lean checks reveal missed translation defects, ambiguous conventions, and incorrect model objections. LLM judgments also help identify formalizations on which human reviewers disagree, without using human ratings to select them. Together, these findings show how LLM judgments can guide review, while source comparisons and targeted checks distinguish translation defects from differences in interpretation and grading criteria.

1 INTRODUCTION

A Lean formalization can compile and still change the mathematics it is meant to express. Translating a natural-language (NL) statement into Lean 4 (de Moura & Ullrich, 2021) may drop a hypothesis, change a variable's domain, or strengthen the conclusion. We assess *faithfulness*: whether the formal statement preserves the mathematical meaning of its source.

LLMs are increasingly used to assess faithfulness (Ying et al., 2024; Gao et al., 2025; Lu et al., 2025; Yu et al., 2025; Zhang et al., 2025; Peng et al., 2025), with their judgments evaluated against human ratings or published benchmark labels. These judgments can disagree for different reasons. An evaluator may miss a translation defect, interpret an ambiguous source differently, or apply a different grading convention. In Figure 1, two human reviewers and one LLM accepted a formalization with a stronger hypothesis than its source, while the other judges rejected it. Published labels can also miss translation defects, as shown in Figure 2. Understanding a disagreement therefore requires examining the source statement, the Lean code, and the criteria behind the judgments.

We investigate these disagreements on benchmarks with published labels and on formalizations with existing human ratings. Comparing model consensus with labels identifies conflicts for targeted checks; rankings constructed from LLM outputs alone help choose which items to inspect before consulting human ratings or labels. We compare individual judgments, simple voting and grade-based rules, and Dawid–Skene (DS) aggregation (Dawid & Skene, 1979), which models observer error rates. We ask three questions:

- **(Q1)** How do individual judges and aggregation rules compare with published labels? Does modelling observer error rates improve verdicts or probability estimates over majority voting?
- **(Q2)** Can LLM judgments alone help choose what to check first: disagreements with benchmark labels, disagreements among human reviewers, or earlier screening rejections that reviewers accepted? How do rankings based on votes, grades, and DS compare?
- **(Q3)** What do targeted Lean checks and source comparisons reveal about the reasons for disagreement? Which cases involve translation defects or mistaken judgments, and which turn on source interpretation or grading conventions?

Our contributions are:

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.

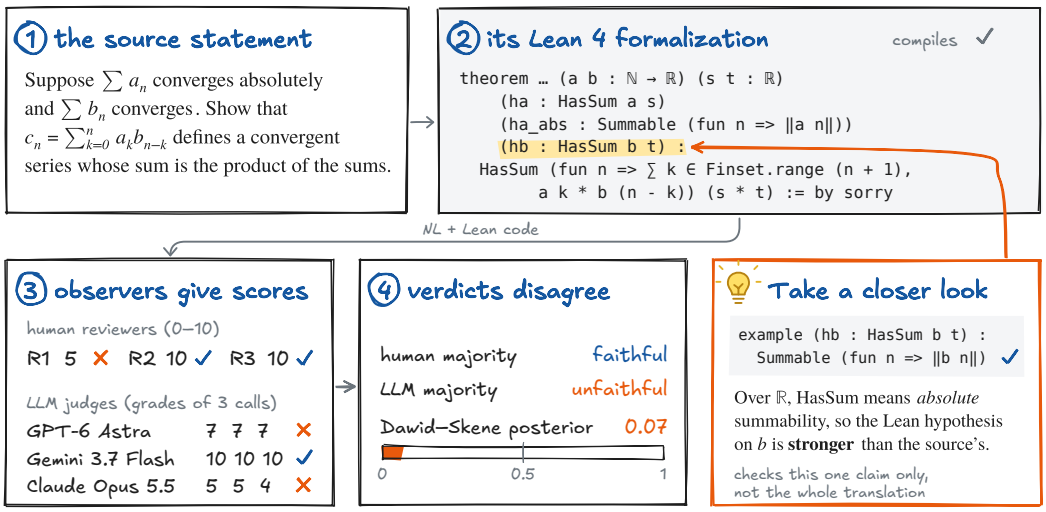


Figure 1: **Two reviewers accepted a formalization whose hypothesis is stronger than its source's.** One human-reviewed item: the natural-language source, its Lean formalization, the grades of three human reviewers and three LLM judges, and the verdicts that different rules reach from the same grades. A grade of at least 9 counts as faithful, and each LLM's vote is the majority of its three queries. A Lean check confirms that `HasSum` over $\mathbb{R}$ requires absolute summability, whereas the source assumes only convergence. The DS posterior shown combines all six observers.

(i) An investigation of disagreements with benchmark labels and human reviewers. We audit 75 CriticLeanBench items, including all 51 on which the three judges unanimously contradict the label. Our audit sides with the judges on 44 of these 51, with the label on 2, and leaves 5 uncertain. Case studies across both settings identify concrete translation defects, ambiguities in source preparation, and differences in grading conventions. Targeted Lean checks provide a counterexample to a statement labeled faithful, expose a strengthened hypothesis accepted by two reviewers, and refute an incorrect LLM objection.

(ii) An evaluation of LLM-only rankings for locating disagreements and revisiting earlier screening rejections. On the human-rated panels, at a 20% review budget, grade margin selects lists in which 77% and 54% of items divide the reviewers, versus 44% and 28% overall; DS uncertainty gives 72% and 45%. Among 97 eligible earlier screening rejections, rankings based on mean grade, acceptance votes, or DS each select 11 of the 18 human-majority acceptances in a 19-item list, versus 3.5 expected at random.

(iii) A comparison of what aggregation adds to individual judgments and simple rankings. DS gives the same verdicts as majority voting on both labeled benchmarks, but its probabilities are over-confident on CriticLeanBench. Its ranking has the highest expected disagreement yield at the smallest tested budget (5%) on both benchmarks, while grade margin leads at larger budgets on CriticLeanBench and at most tested budgets on the human panels. The comparisons distinguish gains in inspection priority from gains in verdicts or confidence.

LLM judgments thus help locate cases worth reviewing, while source comparisons and targeted checks distinguish mathematical defects from differences in interpretation and evaluation criteria.

2 Related Work

Evaluating statement autoformalization. Most Lean benchmarks evaluate proof search against supplied statements (Zheng et al., 2022; Yang et al., 2023; Ren et al., 2025; Wang et al., 2025; Lin et al., 2025). Evaluating generated statements instead uses reference formalizations (Azerbayev et al., 2023), compilation plus back-translation and human examination (Ying et al., 2024; Gao et al., 2025), learned alignment scores (Lu et al., 2025), LLM-based scoring such as LeanScorer (Yu et al., 2025), ensembles of LLM judges (Zhang et al., 2025), critic models trained on human-labeled data (Peng et al., 2025), and one-sided formal equivalence checks (Liu et al., 2025). We study how to

evaluate and use these judgments when the reference itself is fallible: whether aggregation improves decisions, whether LLM outputs identify items worth reviewing, and what targeted checks reveal about disagreements.

LLM-as-a-judge. LLM judges are widely used for open-ended evaluation, with known biases and imperfect agreement with humans (Zheng et al., 2023). Aggregating several judges is a common remedy; we test what such aggregation adds in a formal-mathematics setting.

Noisy-observer estimation. DS estimates each rater's error rates and each item's unobserved true label from the raters' votes (Dawid & Skene, 1979); such estimates are possible with several raters, or with groups of items that differ in how often the true label is positive (Hui & Walter, 1980). When raters' errors are related, these estimates of error rates and prevalence become biased, and the problem is often invisible in the votes alone (Vacek, 1985; Albert & Dodd, 2004). Our labeled benchmarks let us measure this association directly.

3 Experiments and Data

We evaluate LLM judgments on two types of data: formalizations with existing human ratings and benchmarks with published labels. Each item pairs a natural-language (NL) statement with a compiling Lean 4 formalization.

3.1 Data

Formalizations with existing human ratings. The benchmark of Zhang et al. (2026) contains 400 graduate-level statements spanning real analysis, complex analysis, topology, and algebra, without reference Lean formalizations. We analyze two sets of generated formalizations, called *panels*:

- **Panel A**: 116 outputs from the tool-augmented agent of Zhang et al. (2026), selected because they failed an earlier screening rule requiring acceptance by both GPT-5.2 and Gemini-2.5-Pro.
- **Panel B**: 92 outputs from Aristotle (Achim et al., 2025), generated from 100 randomly sampled source statements.

Panel A therefore revisits outputs rejected by the earlier screen, while Panel B includes outputs both accepted and rejected by that screen. Both panels have existing ratings from the same three reviewers on a 0–10 scale. We analyze the panels separately because their generating systems and selection procedures differ. Panel construction, reviewer instructions, and sensitivity analyses appear in Sections C and D.

Benchmarks with published labels. We use 450 pairs from **CriticLeanBench** (Peng et al., 2025): 250 labeled faithful and 200 unfaithful. These labels were produced with DeepSeek-R1 assistance and human verification. We also use 100 pairs from **MA-Align** (MathAtlas, 2026): 25 labeled aligned and 75 misaligned. Further benchmark details appear in Section F.

3.2 LLM Judges and Votes

We query GPT-6 Astra, Gemini 3.7 Flash, and Claude Opus 5.5 three times per item. Each call receives the natural-language statement, its Lean formalization, and the compile status, and returns an integer grade from 0 to 10 with a short explanation.

A formalization is *faithful* if it preserves the source's mathematical meaning. The rubric considers definitions and assumptions that affect the claim, while ignoring verbosity and irrelevant extra lemmas. Proof completeness is not assessed, so `by sorry` is allowed. The prompt and request settings appear in Section B.

Grades of 9 or above count as faithful. Each model's item-level vote is the majority of its three calls, provided all three responses are readable; otherwise, its vote is missing. Human ratings use the same cutoff, with missing ratings left unobserved. Human voting rules appear in Section D.

3.3 JUDGMENT AGGREGATION AND INSPECTION RULES

We compare individual model votes, majority voting across the three models, unanimous acceptance, and Dawid–Skene (DS) (Dawid & Skene, 1979). Unanimous acceptance requires all three models to accept. DS fits the binary votes defined above without reference labels. Let v_{ij} be observer j's vote on item i, s_j and c_j its sensitivity and specificity, and π the faithful proportion. Assuming votes are independent conditional on faithfulness, their likelihoods are

$$L_{i,\mathrm{F}} = \prod_{j \in O_i} s_j^{v_{ij}} (1 - s_j)^{1 - v_{ij}}, \qquad L_{i,\mathrm{U}} = \prod_{j \in O_i} (1 - c_j)^{v_{ij}} c_j^{1 - v_{ij}},$$

where O_i contains observers with recorded votes, and F and U denote faithful and unfaithful. After fitting the parameters by expectation–maximization, the item posterior is

$$\widehat{p}_i = \frac{\widehat{\pi}\, \widehat{L}_{i,\mathrm{F}}}{\widehat{\pi}\, \widehat{L}_{i,\mathrm{F}} + (1 - \widehat{\pi}) \widehat{L}_{i,\mathrm{U}}}.$$

DS accepts items with $\widehat{p}_i \geq 0.5$. Full model and fitting details appear in Section A.

For inspection, we compare random selection with three LLM-only rankings. *Model splits* places non-unanimous votes first. *DS uncertainty* places probabilities closest to 0.5 first. *Grade margin* prioritizes items whose grades least support the majority verdict: it uses the lowest of the nine grades for majority acceptances and the highest for majority rejections, measured relative to the midpoint between grades 8 and 9. We select the same number of items per method and average over random ties (Section G). To revisit earlier rejections, we instead rank by decreasing mean grade, model acceptance vote share, or DS faithfulness probability.

3.4 EXPERIMENTAL SETUP

On the published benchmarks, we evaluate verdicts and probabilities against the labels, count majority–label disagreements selected by each ranking, and audit selected conflicts with the labels. On the human panels, we count selected items with reviewer disagreement or human-majority acceptance of an earlier rejection. Rankings use only LLM outputs; labels and human ratings are used afterwards for evaluation. Targeted checks investigate disagreements in both settings. Protocols, audit procedures, and uncertainty estimation appear in Sections E to J.

4 LLM JUDGES AGAINST PUBLISHED LABELS

Published labels provide a common reference for separating three questions: whether aggregation improves verdicts or probabilities (Q1), whether rankings help locate disagreements (Q2), and what checking those disagreements reveals about the reference itself (Q3).

4.1 AGREEMENT WITH PUBLISHED LABELS

Gemini has the fewest label disagreements on CriticLeanBench and Opus the fewest on MA-Align (Table 1), and DS makes the same decision as majority voting on every item of both benchmarks. Unanimous acceptance reduces false acceptances at the cost of more false rejections (33 FP and 45 FN on CriticLeanBench; Table 10).

DS also changes confidence without changing verdicts. On CriticLeanBench, items with posterior above 0.8 have mean posterior 0.999, but only 84% are labeled faithful; its Brier score is worse than the majority vote share. On MA-Align, DS has a lower Brier score and its mean probability is closer to the label proportion. The judges' votes remain strongly associated within published-label groups, which challenges DS's conditional-independence assumption (Figure 7 and Section F). We next examine disagreements with the labels directly.

4.2 WHAT CHECKING THE DISAGREEMENTS REVEALS

On CriticLeanBench, all three judges contradict the label on 51 of 433 items: they reject 18 labeled faithful and accept 33 labeled unfaithful. To address Q3, we checked all 51, together with 24 other

Table 1: **Agreement with published labels**, on items with three valid responses from every judge. FP: accepted but labeled unfaithful; FN: rejected but labeled faithful. Brier: mean squared probability error (lower is better). Full results: Table 10 and Section F.

	CriticLeanBench ($n = 433$)			**MA-Align** ($n = 96$)		
Method	FP	FN	Brier	FP	FN	Brier
GPT	41	33	.163	11	0	.103
Gemini	43	19	.143	20	0	.193
Opus	37	41	.176	5	1	.065
Majority vote	41	30	.144	8	0	.084
DS	41	30	.162	8	0	.073

items sampled from label disagreements and agreements. The judgments are ours and AI-assisted, so they are not an independent reference; where possible we back them with a mathematical argument, a source comparison, or a Lean check (Table 2). Items are identified by audit IDs (CR*nn*, R*nn*); their CriticLeanBench test IDs appear in Section K.

On 44 of the 51 unanimous disagreements our judgment sides with the judges, on 2 with the label, and 5 remain uncertain (Section H). The strength of the evidence differs by direction. Where the label says faithful and the judges reject, the disagreement usually comes with a concrete, checkable issue (Table 2). Where the label says unfaithful and the judges accept, our acceptances are judgments, not proofs. Checks can also support a label: in CR06, an explicit counterexample to the formal statement supports its negative label and corrected our own first judgment (Section H).

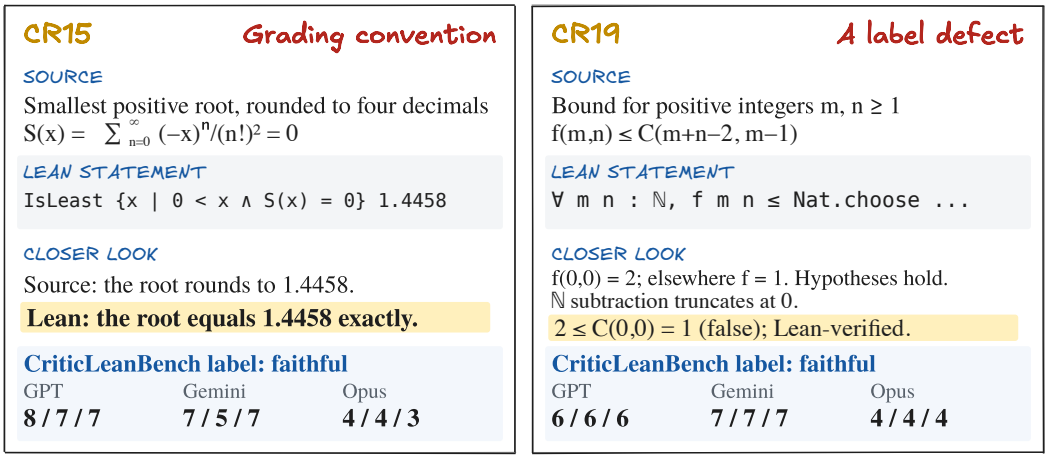


Figure 2: **A grading convention and a label defect.** Both CriticLeanBench items are labeled faithful but rejected by all three LLM judges. Numbers give each model's three grades; acceptance requires at least 9. CR15 contrasts a rounded answer with an exact-root claim. In CR19, the formal conclusion also covers index 0, where a Lean-verified counterexample shows it fails. Statements are abbreviated; details appear in Section K.

The checks explain what a disagreement means. Some disagreements turn on conventions that the binary label leaves implicit, such as whether a rounded answer may be stated as an exact equality (CR15, Figure 2). Others require tracing an item back to its source, since the released text may differ from the original problem (R24). Replacing audited labels with our judgments improves measured agreement and probability error for both DS and majority voting. Their verdicts remain identical, and DS still has the higher Brier score (Q1; Section H.1).

4.3 Choosing what to check without seeing the labels

We test whether LLM-only uncertainty rankings concentrate disagreements with published labels (Q2). We rank items using only LLM outputs (Section 3.4) and reveal the labels only afterwards.

Table 2: **What the checks found**: examples from the 75 audited CriticLeanBench items. R24, R26 and R31 remain uncertain in our audit. Cases and verification details: Sections H and K.

Issue	Example	Evidence
Domain enlargement	CR19 extends positive indices to zero.	Lean-verified counterexample.
Impossible hypothesis	R15 encodes componentwise multiplication instead of orthogonality.	Lean proves the hypothesis impossible.
Division type	R14 and R31 (uncertain) use integer division where the source uses real-valued division.	Lean checks the division types.
Rounding convention	CR15 and R26 (uncertain) replace rounded answers with exact equalities.	Source comparison; numerical or mathematical argument.
Rewritten task	R24 (uncertain) adds a numerator–denominator task absent from the original question.	Comparison with the original contest problem.

At the smallest tested review budget (5%), DS has the highest expected yield of label disagreements among the evaluated rankings on both benchmarks (Table 3). On CriticLeanBench, its 22-item list contains 13.9 of the 71 majority–label disagreements, compared with 11.8 for grade margin, 10.7 for model splits, and 3.6 at random. On MA-Align, its 5-item list contains 3 of the 8 disagreements, versus 1.0 for grade margin, 0.8 for model splits, and 0.4 at random. At 10%, DS remains highest on MA-Align, while the rankings are similar on CriticLeanBench (about 20 of 71, versus 7 at random). At the 20% and 30% budgets, grade margin selects more disagreements on CriticLeanBench; it also leads at 30% on MA-Align (Figure 8, Table 11, and Section G).

Table 3: **Published-label disagreements selected at matched review budgets.** Entries are expected numbers of selected items whose LLM majority verdict disagrees with the published label, averaging over random ties. All rankings use only LLM outputs. Bold marks the highest expected count in each column, including ties. Full budget curves and protocols: Figure 8 and Section G.

	CriticLeanBench ($n = 433$)				**MA-Align** ($n = 96$)			
Ranking	5%	10%	20%	30%	5%	10%	20%	30%
Items reviewed	22	43	87	130	5	10	19	29
Random selection	3.6	7.1	14.3	21.3	0.4	0.8	1.6	2.4
Model splits first	10.7	20.3	26.0	31.6	0.8	1.6	**3.0**	3.6
DS uncertainty	**13.9**	20.2	25.2	30.1	**3.0**	**3.0**	**3.0**	4.8
Grade margin	11.8	**21.2**	**30.9**	**40.0**	1.0	1.0	2.9	**5.2**

The methods use different information. DS distinguishes voting patterns that model splits treats alike, but assigns the same probability to all items sharing a pattern. Grade margin retains differences in their grades (Table 13 and Section G). This matters because 51 of the 71 CriticLeanBench label disagreements occur where all three models agree. Comparing consensus with the labels and ranking items without labels therefore provide complementary ways to select cases for review.

5 LLM AND HUMAN JUDGMENTS WITHOUT GROUND TRUTH

We next examine the human-rated panels, where no reference label is trusted. We ask whether LLM judgments help locate reviewer disagreements and earlier screening rejections supported by a human majority (Q2), and what Lean checks reveal about individual disagreements (Q3).

5.1 HUMANS AND LLMS DISAGREE, AND NEITHER IS RELIABLE BY DEFAULT

Acceptance rates have different interpretations across the panels (Figure 3): on Panel A they show how often observers overturn earlier rejections; on Panel B they summarize judgments of Aristotle's outputs. On Panel A, R3 accepts 9 of 90 rated items, compared with 44 of 114 for R1.

Among the LLMs, Gemini accepts most and Opus least on both panels. The reviewers agree only moderately with one another: Cohen's κ is 0.15–0.27 on Panel A and 0.31–0.41 on Panel B. R1 is

the least accepting reviewer on B and the most accepting on A, showing that acceptance behavior depends on the item set.

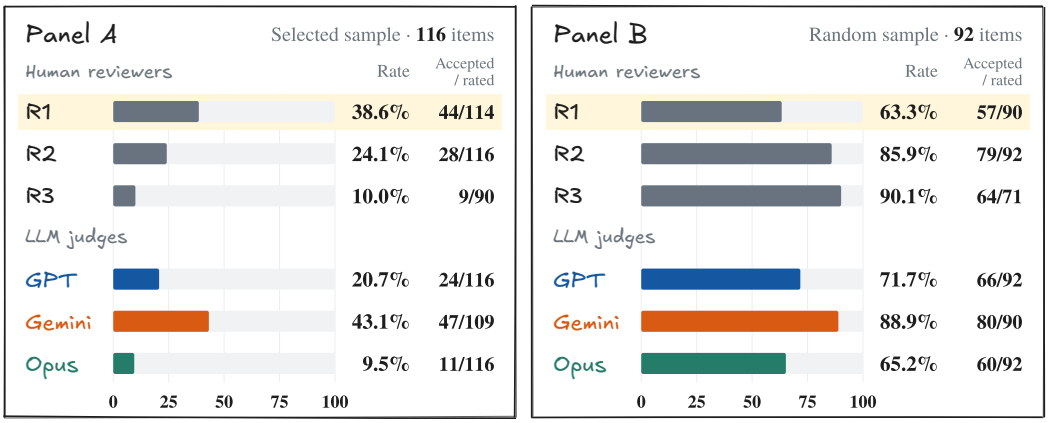


Figure 3: **Acceptance rates in two screening contexts.** Bars show the share of rated items judged faithful (grade $\geq$ 9); counts are accepted/valid ratings. Each model's vote is the majority of its three calls. **Panel A:** all outputs were rejected by the earlier screen of Zhang et al. (2026), so lower acceptance means greater agreement with those rejections. **Panel B:** Aristotle's outputs include both earlier acceptances and rejections; higher acceptance means that the observer judges more of these formalizations faithful.

Neither side can serve as truth by default. In B081 (Figure 1), two reviewers accepted a formalization whose Lean hypothesis requires absolute summability where the source requires only convergence; GPT, Opus, and R1 rejected it. Conversely, on B099, which all three reviewers accepted, a Lean check refutes a GPT objection about summation notation (Section E).

5.2 LLM-ONLY RANKINGS SELECT ITEMS THAT DIVIDE REVIEWERS

We compute the three LLM-only rankings described in Section 3.3 on each panel, fitting DS to the LLM votes alone; no human grade enters any ranking. Among items with all three LLM votes and at least two human grades, we then count how many selected items have *human disagreement*: at least one grade of 9 or more and another below 9. We use the same 5%, 10%, 20%, and 30% budgets in both settings, adding the 5% comparison to the panel analysis with the fitted rankings held fixed.

Table 4: **Human disagreements selected at matched review budgets.** Entries are expected numbers of selected items on which human reviewers disagree, averaging over random ties. All rankings use only LLM outputs. Bold marks the highest expected count in each column, including ties. Intervals at the 20% budget: Table 17.

	Panel A ($n = 107$)				**Panel B** ($n = 89$)			
Ranking	5%	10%	20%	30%	5%	10%	20%	30%
Items reviewed	5	11	21	32	4	9	18	27
Random selection	2.2	4.8	9.2	14.1	1.1	2.5	5.1	7.6
Model splits first	3.7	8.1	15.5	23.6	**1.9**	4.3	8.7	**13.0**
DS uncertainty	3.6	8.0	15.1	23.4	1.2	2.7	8.1	**13.0**
Grade margin	**5.0**	**9.1**	**16.2**	**24.0**	1.6	**4.7**	**9.7**	12.7

At a 20% review budget, grade margin selects an expected 16.2 items with human disagreement out of 21 on Panel A (77.2%) and 9.7 out of 18 on Panel B (53.7%), compared with 9.2 (43.9%) and 5.1 (28.1%) at random. Model splits and DS also concentrate disagreements: their expected counts are 15.5 and 15.1 on Panel A, and 8.7 and 8.1 on Panel B (Table 4). Grade margin has the highest point estimate at every tested budget on Panel A; the leading method varies with the budget on Panel B. The smaller panel has wider uncertainty, and the 20% gain over random selection includes zero for

DS (Section I). These rankings help locate disputed judgments using LLM outputs alone. In B081, both the models and reviewers split, and a targeted Lean check identifies a strengthened hypothesis.

5.3 Revisiting earlier screening rejections

Panel A also lets us ask which earlier rejections deserve reconsideration. We rank by decreasing mean grade, model acceptance vote share, or LLM-only DS faithfulness probability, and count items accepted by at least two human reviewers. These rankings prioritize likely acceptances rather than disputed judgments. Requiring complete model votes and a decisive human majority leaves 97 items, of which 18 are accepted by that majority.

Each ranking finds 11 of these 18 candidates (61%) in its 19-item list, compared with 3.5 expected at random. Human-majority acceptance is 58% in each selected list, versus 19% in the eligible pool. LLM judgments thus help direct reconsideration of earlier rejections, with the same yield from simple vote and grade rankings as from DS at this budget. Protocols, intervals, and the corresponding analysis on Panel B appear in Section J.

Combining the collected judgments. A supplementary DS analysis treats human ratings as inputs alongside LLM votes. Adding human judgments changes the estimated faithful share from 18% to 31% on Panel A and from 76% to 73% on Panel B. Observer parameters and sensitivity analyses are reported in Section D.1; these fits describe how aggregation changes when human judgments are included.

6 Discussion

LLM judgments can help examine the references used to evaluate formalizations. On CriticLeanBench, comparing unanimous model judgments with published labels identifies conflicts whose audit supports the models in 44 of 51 cases. On the human panels, rankings built without human ratings concentrate reviewer disagreements and identify earlier screening rejections supported by a human majority. These are complementary uses: consensus comparison draws on an existing reference, while the rankings choose items from LLM outputs alone.

Simple methods account for much of this usefulness. Grade margin leads at most tested budgets on the human panels and at larger budgets on CriticLeanBench; DS leads at the smallest tested benchmark budget. For reconsidering earlier rejections, mean grade, acceptance votes, and DS select the same number of human-supported candidates at a 20% budget. Choosing an inspection order therefore warrants a direct comparison of these methods, rather than assuming that a more elaborate aggregation model will improve it. DS's unchanged verdicts and overconfident probabilities on CriticLeanBench also show why verdicts, confidence, and inspection priority need separate evaluation.

Targeted checks explain the selected conflicts. They reveal a missed domain restriction (CR19), a grading convention left implicit by a label (CR15), benchmark text rewritten from its source (R24), a strengthened hypothesis two reviewers missed (B081), and an incorrect model objection (B099). The evidence can support the judges, the reference, or an unresolved interpretation (Table 16).

In practice, these results support retaining judges' grades, comparing their consensus with available labels, and evaluating inspection rankings at equal review budgets. Reviewing unanimous judgments as well as split votes is important: most CriticLeanBench majority–label disagreements occur where the models agree.

7 Limitations

The two panels differ in selection, content and generating system, so their rates cannot rank systems. Some items lack full human ratings, and thirty Panel-B items served as a development pilot. Published labels may be wrong, and proprietary judges may have seen benchmark items during training. The ranking analyses are exploratory; their panel outcomes are observed reviewer disagreement or majority acceptance, rather than verified faithfulness. They do not measure whether subsequent

reviewers correct the selected items. Our audit is AI-assisted and not an independent validation; because it targets disagreements, it cannot estimate how often benchmark labels are wrong.

8 Conclusion

LLM judgments can guide the review of both benchmark labels and human assessments of Lean formalizations. Auditing conflicts between model consensus and published labels reveals concrete label problems. Rankings built from LLM outputs alone concentrate human disagreements and identify earlier screening rejections supported by reviewers. Simple vote and grade rules are strong baselines, while the value of DS depends on the dataset and review budget. Together, these results support using LLM judgments to choose where to look and targeted checks to explain what is found.

References

Tudor Achim, Alex Best, Alberto Bietti, Kevin Der, Mathïs Fédérico, Sergei Gukov, Daniel Halpern-Leistner, Kirsten Henningsgard, Yury Kudryashov, Alexander Meiburg, Martin Michelsen, Riley Patterson, Eric Rodriguez, Laura Scharff, Vikram Shanker, Vladmir Sicca, Hari Sowrirajan, Aidan Swope, Matyas Tamas, Vlad Tenev, Jonathan Thomm, Harold Williams, and Lawrence Wu. Aristotle: IMO-level automated theorem proving. *arXiv preprint arXiv:2510.01346*, 2025. URL https://arxiv.org/abs/2510.01346.

Paul S. Albert and Lori E. Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, 2004.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language (system description). In *Automated Deduction – CADE 28*, volume 12699 of *LNCS*, pp. 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

Stephen R. Doty. Lecture notes on abstract algebra. https://github.com/srdoty/AbstractAlgebraBook, 2025. GitHub repository; accessed 2025-09-04.

Guoxiong Gao, Yutong Wang, Jiedong Jiang, Qi Gao, Zihan Qin, Tianyi Xu, and Bin Dong. Herald: A natural language annotated lean 4 dataset. In *The Thirteenth International Conference on Learning Representations*, 2025.

S. L. Hui and S. D. Walter. Estimating the error rates of diagnostic tests. *Biometrics*, 36(1):167–171, 1980.

Jiří Lebl. Basic analysis: Introduction to real analysis. https://github.com/jirilebl/ra, 2025a. GitHub repository; accessed 2025-09-04.

Jiří Lebl. Guide to cultivating complex analysis: Working the complex field. https://github.com/jirilebl/ca, 2025b. GitHub repository; accessed 2025-09-04.

Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, et al. Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction. *arXiv preprint arXiv:2508.03613*, 2025. URL https://arxiv.org/abs/2508.03613.

Qi Liu, Xinhao Zheng, Xudong Lu, Qinxiang Cao, and Junchi Yan. Rethinking and improving autoformalization: Towards a faithful metric and a dependency retrieval-based approach. In *Proceedings of ICLR 2025 (Spotlight)*, 2025. URL https://openreview.net/forum?id=hUb2At2DsQ.

Jianqiao Lu, Yingjia Wan, Yinya Huang, Jing Xiong, Zhengying Liu, and Zhijiang Guo. Formalalign: Automated alignment evaluation for autoformalization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=B5RrIFMqbe.

MathAtlas. MA-Align. https://huggingface.co/datasets/MathAtlas/MA-Align, 2026. Dataset revision 537e9a696247e38ff0e1ba2ed1330490b0be1840. TODO: replace with the paper citation.

Benjamin McKay. Topology lecture notes. https://github.com/Ben-McKay/topology-lecture-notes, 2025. GitHub repository; accessed 2025-09-04.

Zhongyuan Peng, Yifan Yao, Kaijing Ma, Shuyue Guo, Yizhe Li, Yichi Zhang, Chenchen Zhang, Yifan Zhang, Zhouliang Yu, Luming Li, Minghao Liu, Yihang Xia, Jiawei Shen, Yuchen Wu, Yixin Cao, Zhaoxiang Zhang, Wenhao Huang, Jiaheng Liu, and Ge Zhang. Criticlean: Critic-guided reinforcement learning for mathematical formalization. *arXiv preprint arXiv:2507.06181*, 2025. doi: 10.48550/arXiv.2507.06181.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

Pamela M. Vacek. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41(4):959–968, 1985.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, et al. Kimina-prover preview: Towards large formal reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.11354*, 2025.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Huaiyuan Ying, Zijian Wu, Yihan Geng, Jiayu Wang, Dahua Lin, and Kai Chen. Lean workbook: A large-scale lean problem set formalized from natural language math problems. *Advances in Neural Information Processing Systems*, 37:105848–105863, 2024.

Xuejun Yu, Jianyuan Zhong, Zijin Feng, Pengyi Zhai, Roozbeh Yousefzadeh, Wei Chong Ng, Haoxiong Liu, Ziyi Shou, Jing Xiong, Yudong Zhou, et al. Mathesis: Towards formal theorem proving from natural languages. *arXiv preprint arXiv:2506.07047*, 2025. URL https://arxiv.org/abs/2506.07047.

Ke Zhang, Patricio Gallardo Candela, Sudhir Murthy, Yi Xie, Zhi Wang, and Maziar Raissi. Beyond compilation: Evaluating faithful natural-language-to-lean statement formalization. *arXiv preprint arXiv:2606.31002*, 2026.

Lan Zhang, Marco Valentino, Jordan Meadows, and Andre Freitas. Beyond gold standards: Epistemic ensemble of LLM judges for formal mathematical reasoning. *arXiv preprint arXiv:2506.10903*, 2025. URL https://arxiv.org/abs/2506.10903.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. minif2f: a cross-system benchmark for formal olympiad-level mathematics. In *ICLR*, 2022.

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 (Datasets and Benchmarks Track)*, 2023.

A Dawid–Skene Model and Fitting

Model and assumptions. Each LLM contributes one binary vote per item, obtained by majority voting over its three calls. Each human reviewer contributes one vote by thresholding their score at 9. An observer therefore denotes an LLM model or a human reviewer.

Let $z_i \in \{\mathrm{F}, \mathrm{U}\}$ denote the latent faithfulness of item i, and let $v_{ij} = 1$ if observer j accepts it and $v_{ij} = 0$ otherwise. Let O_i contain the observers with a recorded vote, and let $\pi = P(z_i = \mathrm{F})$ be the faithful proportion. For each observer, DS estimates sensitivity and specificity:

$$s_j = P(v_{ij} = 1 \mid z_i = \mathrm{F}), \qquad c_j = P(v_{ij} = 0 \mid z_i = \mathrm{U}).$$

DS assumes that observers' votes are independent conditional on an item's latent status. The conditional likelihoods of its observed votes are therefore

$$L_{i,\mathrm{F}} = \prod_{j \in O_i} s_j^{v_{ij}} (1 - s_j)^{1 - v_{ij}}, \qquad L_{i,\mathrm{U}} = \prod_{j \in O_i} (1 - c_j)^{v_{ij}} c_j^{1 - v_{ij}},$$

and the item likelihood is

$$P(v_i) = \pi L_{i,\mathrm{F}} + (1 - \pi) L_{i,\mathrm{U}}.$$

Missing votes contribute no factor to these products.

Estimation and outputs. We estimate π, s_j, and c_j by expectation–maximization from multiple starting points, retaining the highest-likelihood fit. Class orientation, parameter bounds and fitting details appear in Section D.

At the fitted parameter values, the posterior probability of faithfulness is

$$\widehat{p}_i = P(z_i = \mathrm{F} \mid v_i; \widehat{\theta}) = \frac{\widehat{\pi}\, \widehat{L}_{i,\mathrm{F}}}{\widehat{\pi}\, \widehat{L}_{i,\mathrm{F}} + (1 - \widehat{\pi}) \widehat{L}_{i,\mathrm{U}}},$$

where $\widehat{\theta}$ denotes the fitted parameters. The posterior $\widehat{p}_i$ serves three purposes: thresholding it at 0.5 gives a binary verdict, its value expresses the model's estimated probability of faithfulness, and it determines inspection priority according to the review task. Items closer to 0.5 are inspected first when seeking disputed judgments; items with the highest probabilities are inspected first when reconsidering earlier screening rejections (Section J). We also examine the estimated faithful proportion $\widehat{\pi}$, which differs from the fraction of items receiving a faithful verdict.

LLM-only fits determine inspection priorities, while combined human–LLM fits examine the stability of estimates across observer sets. Evaluation procedures and sensitivity analyses appear in Section 3.4.

Uncertainty in panel estimates. Intervals for the DS estimates on the panels use an item bootstrap: items are resampled with replacement and DS is refitted on each replicate. These intervals reflect sampling variability of the items but not uncertainty in the model specification. Benchmark comparisons and inspection rankings instead use paired resampling of NL groups, keeping items sharing a source statement together; their procedures appear in Sections F, G, I and J.

B Judge Rubric and Request Settings

The following prompt was used for all model judgments. Grading thresholds are applied outside the prompt.

Table 5: Model request settings.

Model	Request configuration
GPT-6 Astra	OpenAI API; medium reasoning; JSON output; maximum 32,768 output tokens.
Gemini 3.7 Flash	Google API; medium thinking; temperature 1.0; JSON MIME type; maximum 32,768 output tokens.
Claude Opus 5.5	Anthropic API; medium effort; maximum 32,768 output tokens.

```
You are an expert in Lean 4, Mathlib, and mathematics. You are judging TRANSLATION-ONLY: the
    formalization of a natural-language statement into Lean 4.

Input: (1) a natural-language statement, (2) a Lean 4 code snippet, (3) a compile_pass boolean
    .
Your job: grade whether the Lean statement expresses the same mathematical meaning as the
    natural-language statement. You are NOT evaluating a proof.

Key policy:
- If compile_pass = False: grade must be 0..3.
- If compile_pass = True: ignore proof bodies entirely, including `by sorry`. Proof
    correctness, proof completeness, and the use of helper lemmas inside proofs are NOT part
    of this evaluation.
- Definition bodies that determine the meaning of the target statement are part of the
    statement-level evaluation. Do not ignore them as if they were proofs.
- Do not deduct points for verbosity, extra declarations, or failure to output exactly one
    declaration alone. Grade semantic faithfulness, not output style or prompt compliance.

Identify the target statement:
- Identify the declaration corresponding to the natural-language claim or definition. It may
    be a theorem, lemma, example, or definition; do not assume it is the last declaration in
    the snippet.
- Evaluate its full statement, including relevant surrounding variables, implicit parameters,
    typeclass assumptions, local notation, and definitions.
- Extra declarations do not affect the grade unless they affect the meaning of the target
    statement.
- If the code only states partial components and never expresses the full natural-language
    claim, treat the missing content as a semantic mismatch. Do not infer an unstated target
    claim from helper lemmas.

Helper definitions and lemmas:
- Helper definitions and abbreviations are allowed when they preserve the intended meaning.
- Inspect helpers needed to interpret the target statement, following their definitions as
    necessary. Do not assume a helper encodes the intended concept merely because its name
    sounds appropriate.
- Ignore lemmas used only in proofs, whether or not their proofs are complete. Unused or
    additional lemmas do not justify a deduction by themselves.
- Penalize a placeholder or altered definition when it replaces an essential mathematical
    concept with a different or unspecified meaning. Judge the actual semantic effect, not a
    keyword alone.
- Do not automatically penalize `True`, `False`, `0`, `none`, `default`, `Classical.choice`, `
    axiom`, `opaque`, or `sorry` merely for appearing. In particular, a proof ending in `by
    sorry` is acceptable. Distinguish a missing proof from missing definition content needed
    to specify the intended concept.
- Check additional hypotheses, including helper propositions or a version of the conclusion
    introduced as an assumption. Penalize them when they are not stated or implied by the
    natural-language statement and materially change the claim.

How to judge meaning:
- Compare the target statement, interpreted with its relevant definitions and context, to the
    natural-language statement.
- Check quantifiers, their scope and dependencies, and logical structure: forall/exists,
    implication, equivalence, conjunction, and disjunction.
- Check hypotheses, conclusions, and main objects/domains: natural numbers, integers, real or
    complex numbers, rings/groups, ZMod n, matrices, and other relevant structures.
- Check for omitted conditions, additional restrictions, and strengthened or weakened claims.
- Accept mathematically equivalent formulations and harmless implementation or
    representational differences when the intended meaning is preserved.

Scoring guide (integer 0..10):
- 0-3: compile_pass is False, or the statement is unrelated or fundamentally wrong. Use 0 for
    an unrelated translation; when compilation fails, keep the grade within 0..3.
- 4-6: compiles, but meaning differs substantially, for example through a wrong domain, a
    missing essential condition or conclusion, or an essential concept defined incorrectly.
- 7-8: compiles and captures much of the intended statement, but still has a meaningful
    semantic mismatch, such as an unjustified strengthening or weakening.
- 9: compiles and preserves the intended mathematical meaning, with only harmless
    representational differences. Do not assign 9 if a meaningful semantic mismatch remains.
- 10: compiles and fully matches the intended mathematical meaning.

Output contract (STRICT):
Return a single JSON object with exactly two fields:
- "grade": an integer from 0 to 10.
- "thought": a short explanation focusing on the statement-level comparison, enclosed by the
    markers "### BEGIN THOUGHT" and "### END THOUGHT" with newline separators.

Example output format:
{
  "grade": 10,
  "thought": "### BEGIN THOUGHT\n<short explanation focusing on statement-level comparison>\n
    ### END THOUGHT"
```

```
}

Return ONLY valid JSON. Do not return a "faithful" field, extra keys, or markdown outside the
    JSON.
```

C Human Review Protocol

Reviewers scored each case on a 0–10 scale (0: unrelated or non-compiling; 1–3: clearly wrong; 4–6: related but materially different; 7–8: mostly matches with a noticeable strengthening, weakening, or domain mismatch; 9: very close; 10: faithful), ignoring proofs and `by sorry`, allowing faithful helper definitions, and penalizing vacuous placeholder definitions. Reviewers saw only the NL statement, the Lean code, and its compile status. Coverage is R1 114/116, R2 116/116, R3 90/116 on A, and 90/92, 92/92, 71/92 on B. All 26 A and 21 B abstentions by R3 carry a "Problematic NL" note.

D Additional Panel Results

Table 6: Source benchmark composition.

Domain	**Source material**	**N**
Real Analysis	*Basic Analysis* (Lebl, 2025a)	100
Complex Analysis	*Cultivating Complex Analysis* (Lebl, 2025b)	100
Topology	*Notes on Topology* (McKay, 2025)	100
Algebra	*Abstract Algebra* (Doty, 2025)	100
Total		**400**

Table 7: **Acceptance (grade ≥ 9) by observer.** Model votes are three-call majorities. Denominators exclude missing ratings; Panel A is a selected subset (Section 3.1).

Observer	**Panel B (random, $n = 92$)**	**Panel A (selected, $n = 116$)**
R1	57/90 (63%)	44/114 (39%)
R2	79/92 (86%)	28/116 (24%)
R3	64/71 (90%)	9/90 (10%)
GPT-6 Astra	66/92 (72%)	24/116 (21%)
Gemini 3.7 Flash	80/90 (89%)	47/109 (43%)
Claude Opus 5.5	60/92 (65%)	11/116 (9%)
Human majority ($\geq$2 agreeing)	69/84	19/104 (12 undecided)

If one reviewer accepts an item, one rejects it, and the third did not score it, the human majority is undecided.

Observer agreement. The repeated LLM calls are stable: on Panel B the three calls split on 2 GPT, 1 Gemini, and 7 Opus items (of 92, 90, and 92 with three valid calls); on Panel A they split on 4, 6, and 3 (of 116, 109, and 116). Cohen's κ compares observed agreement with the agreement expected from the two reviewers' acceptance rates (1: perfect; 0: chance). On Panel B, R1–R2 agree on 65/90 items ($\kappa = 0.31$), R1–R3 on 56/70 (0.41), and R2–R3 on 65/71 (0.36). On Panel A, κ is 0.25, 0.15, and 0.27 for these pairs. Removing R1's deductions for unused helper lemmas raises R1–R2 agreement on B from 65/90 to 75/90, but also restores a score of 10 on an item whose bound drops a factor of $\frac{1}{2}$ (B026, Section K).

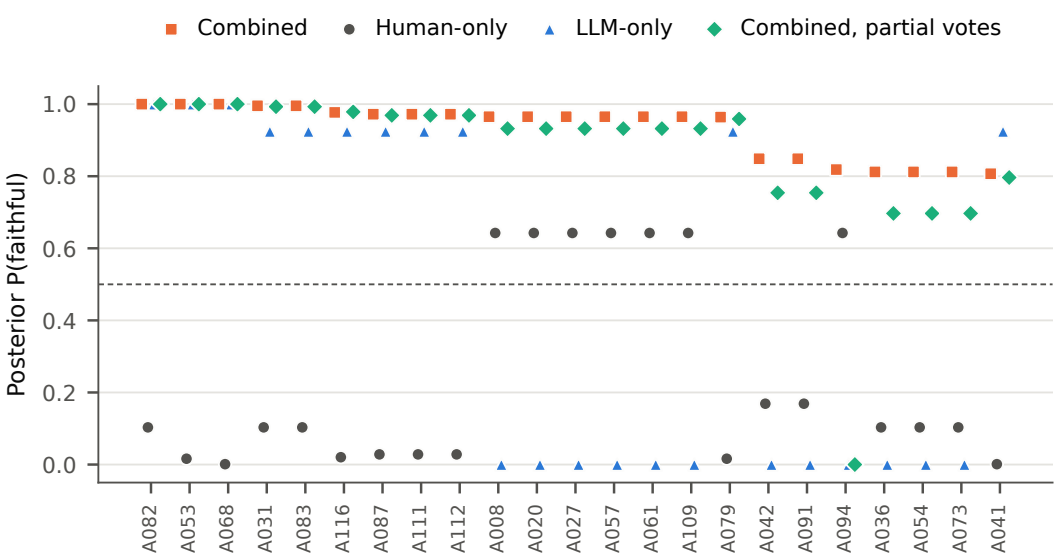


Figure 4: **Panel-A items whose DS label changes with the observer subset or missing-call rule.** Human-only and LLM-only fits disagree on many items; under the complete-triple rule A094 receives posterior 0.82, while counting Gemini's two valid rejecting calls moves it to near 0.

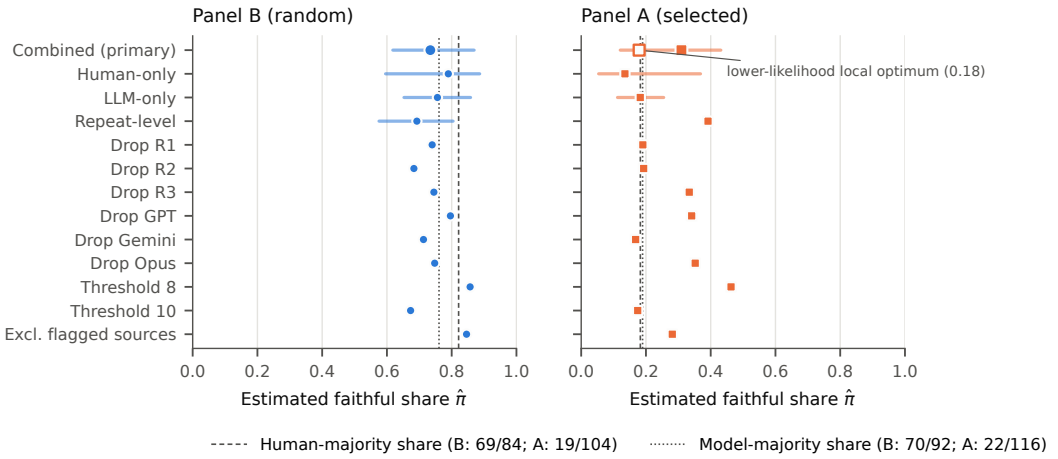


Figure 5: **Sensitivity of the DS faithful-share estimate** to panel composition, repeat handling, observer removal, threshold, and exclusion of reviewer-flagged sources. Horizontal bars are item-bootstrap 95% intervals, computed only for the first four specifications (and not for repeat-level on Panel A); other rows are point estimates. Vertical lines mark voting shares. On Panel A the estimate moves between 0.14 and 0.46, and the combined model has a lower-likelihood local optimum (open marker).

DS fitting. We fit DS by expectation–maximization from the majority vote and several random starts, keeping the highest-likelihood fit. Parameters are kept within $[10^{-6}, 1 - 10^{-6}]$, and the classes are oriented so that the observers' mean of $s_j + c_j - 1$ is positive. The repeat-level variant treats each LLM query as a vote, with the three queries of one model sharing a sensitivity and a specificity.

Model fit. We simulated votes from each fitted DS model, keeping the observed pattern of missing votes, and compared how often pairs of observers agree. Some pairs agree more often in the real data than the model predicts: R2 and Gemini on Panel B (85 of 90 observed, against 77.1 predicted;

90% range 73–81), GPT and Opus on Panel A (103 of 116, against 94.3), and R3 and GPT on Panel A (79 of 90, against 71.2). Fitted parameters also differ between panels; for example, R3's fitted sensitivity is 0.98 on Panel B and 0.25 on Panel A.

Posterior stability. For each bootstrap fit, we evaluate the posterior on every item's observed vote pattern. The central 95% range of these posteriors exceeds 0.5 in width for 21 Panel-B and 37 Panel-A items. On Panel B, only 5 posteriors of the main fit lie between 0.2 and 0.8. R2's fitted sensitivity reaches the upper bound, so every item R2 rejects receives a posterior below 4×10^{-6}, including one accepted by R3, GPT and Gemini. The estimated proportion spans 0.67–0.86 across sensitivity analyses, with the grade threshold having the largest effect. On Panel A, a second local optimum gives $\pi = 0.179$ versus 0.310 for the main fit; changing the observer subset gives estimates from 0.14 to 0.35, and all sensitivity analyses span 0.14–0.46. For one item, counting Gemini's two valid rejections despite an unreadable third response changes the posterior from 0.82 to about 3×10^{-6} (Figure 4).

D.1 Combining human and LLM judgments

The main-text ranking analyses use human ratings only for evaluation. Here we include those ratings as inputs to DS, jointly fitting the three human reviewers and three LLM judges without reference labels. We compare LLM-only, human-only, and combined fits within each panel, keeping the acceptance threshold at 9. On Panel A, π estimates the faithful share among earlier screening rejections; on Panel B, it estimates the share among the sampled compiling outputs.

Table 8: **DS on the two panels.** π is the estimated faithful share. Intervals are item-bootstrap 95% percentile intervals (Section A).

Specification	Panel A	Panel B
Combined (3 humans + 3 models)	**0.310** (0.121–0.431)	**0.734** (0.618–0.869)
second local optimum	0.179	none
Human-only	0.135 (0.054–0.369)	0.790 (0.596–0.887)
LLM-only	0.183 (0.112–0.255)	0.756 (0.653–0.858)
Repeat-level (shared per-model parameters)	0.392	0.693 (0.576–0.804)
Drop one observer (range)	0.168–0.352	0.683–0.796
Threshold 8 / 10	0.463 / 0.175	0.857 / 0.673
Exclude R3-flagged problematic sources	0.281	0.846
Voting baselines		
Human majority	19/104	69/84
Majority of 3 models	22/116	70/92
Majority of all 6 observers	16/105	70/87

Adding human judgments raises the estimated faithful share on Panel A from 18% to 31% (Table 8). The combined estimate also exceeds the human-only estimate of 14%: incorporating both groups jointly re-estimates observer error rates and changes how their votes contribute to inferred faithfulness. On Panel B, the three estimates are closer: 76% from LLMs, 79% from humans, and 73% from the combined fit.

The fitted observer parameters also explain how DS resolves conflicting judgments. On Panel B, R2's sensitivity reaches the upper bound, making R2's rejections nearly decisive: every item R2 rejects receives posterior below 4×10^{-6}, including one accepted by R3, GPT, and Gemini. Combining the judgments thus provides both a panel-level estimate and an account of which observers drive individual verdicts. The remaining rows of Table 8 summarize sensitivity to observer removal, grade thresholds, and other analysis choices; fitting diagnostics and posterior sensitivity appear in Section D.

E LEAN-VERIFIED DIAGNOSTICS

All checks run in Lean 4.28.0 with the Mathlib version used to produce the outputs. They verify specific claims, not complete faithfulness labels. Full source/code/score case studies for B081, B099, and B026 appear in Section K.

- Big-operator scope: `∑ k ∈ s, f k - e` parses as `(∑ k ∈ s, f k) - e` (refutes a GPT objection; B099).
- `HasSum b t` over $\mathbb{R}$ implies `Summable (λ n, ‖b n‖)` (a strengthened hypothesis missed by two reviewers; B081).
- A stated bound $|\cosh z| \le \frac{1}{2}e^{|z|}$ is false at $z = 0$; the formalization weakens it (B026).

F EXTERNAL BENCHMARK DETAILS

Table 9: **DS-fitted vs. label-based rates and label-conditional odds ratios** on the CriticLeanBench common subset (433 items). Odds ratios are conditional maximum-likelihood estimates with exact 95% intervals.

	Sensitivity		Specificity	
Judge	DS	Labels	DS	Labels
GPT-6 Astra	0.976	0.866	0.983	0.781
Gemini 3.7 Flash	0.996	0.923	0.920	0.770
Claude Opus 5.5	0.937	0.833	0.994	0.802
Odds ratio	faithful label ($n = 246$)		unfaithful label ($n = 187$)	
GPT–Gemini	240 (33.8–10,641)		318 (71.3–2,168)	
GPT–Opus	115 (33.5–524)		211 (50.7–1,336)	
Gemini–Opus	∞ (38.8–∞)		283 (58.1–2,835)	

On MA-Align's common subset, all three judges accept nearly every faithful item (odds ratios undefined); among the 72 unfaithful items, odds ratios are 10.4 (2.1–70.2) for GPT–Gemini and unbounded for the other pairs (lower bounds 6.9 and 2.7).

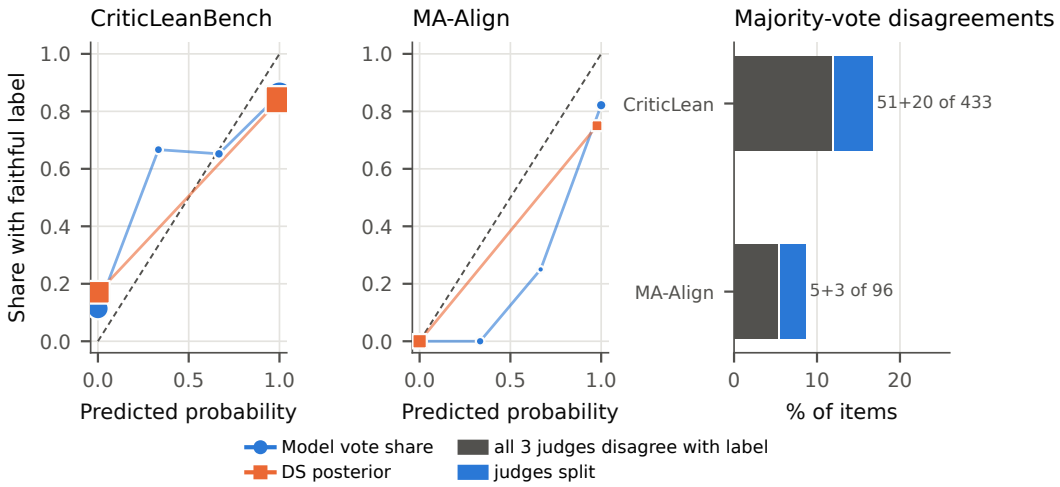


Figure 6: **Calibration and majority-vote disagreements** on the common subset. Left, middle: reliability of the model vote share and the DS posterior in five equal-width bins (empty bins omitted; marker area increases with items per bin on one shared scale; dashed: perfect calibration). DS posteriors occupy only the extreme bins. Right: majority-vote disagreements with the published label, split into items where all three judges agree against the label and items where they split.

F.1 COMMON-SUBSET COMPARISONS AND PROBABILITY DEFINITIONS

For probability-based comparisons, a single judge assigns $p = k/3$ when k of its calls accept. The majority method uses the fraction of model votes accepting; DS uses its fitted posterior, and unanimity uses 0 or 1. Their mean, $\bar{p}$, estimates the share of faithful items.

Table 10: **Comparison with published labels** where every judge returned three valid responses: CriticLeanBench 433/450, MA-Align 96/100. FP: accepted but labeled unfaithful; FN: rejected but labeled faithful. Header denominators give the corresponding label counts. "Accept" is the fraction classified faithful; $\bar{p}$ is the mean predicted probability. Brier is the average squared difference between each item's probability and its label, coded 1 for faithful and 0 otherwise (lower is better).

	CriticLeanBench					MA-Align				
Method	FP/187	FN/246	Accept	$\bar{p}$	Brier	FP/72	FN/24	Accept	$\bar{p}$	Brier
GPT-6 Astra	41	33	.587	.587	.163	11	0	.365	.358	.103
Gemini 3.7 Flash	43	19	.624	.623	.143	20	0	.458	.455	.193
Claude Opus 5.5	37	41	.559	.562	.176	5	1	.292	.299	.065
Majority of 3	41	30	.594	.590	.144	8	0	.333	.372	.084
Unanimous accept	33	45	.540	.540	.180	5	1	.292	.292	.062
DS	41	30	.594	.594	.162	8	0	.333	.326	.073
Label share			.568	.568				.250	.250	

Verdicts and probabilities. DS gives every item the same verdict as the three-model majority on both benchmarks (Table 10), but their probabilities differ. On MA-Align, the mean DS probability is closer to the label share than the mean vote share is (absolute difference 0.076 vs. 0.122; paired difference -0.046, 95% interval -0.083 to -0.014). No difference is detected on CriticLeanBench (0.026 vs. 0.022; interval -0.007 to $+0.015$).

For individual probabilities, DS's Brier score is worse on CriticLeanBench than the majority's ($+0.018$, interval $+0.008$ to $+0.029$); no difference is detected on MA-Align. The best single judge also differs between benchmarks. Unanimous acceptance reduces false acceptances relative to the majority at the cost of more false rejections. Intervals use 1000 paired resamples of NL statements and are not adjusted for multiple comparisons.

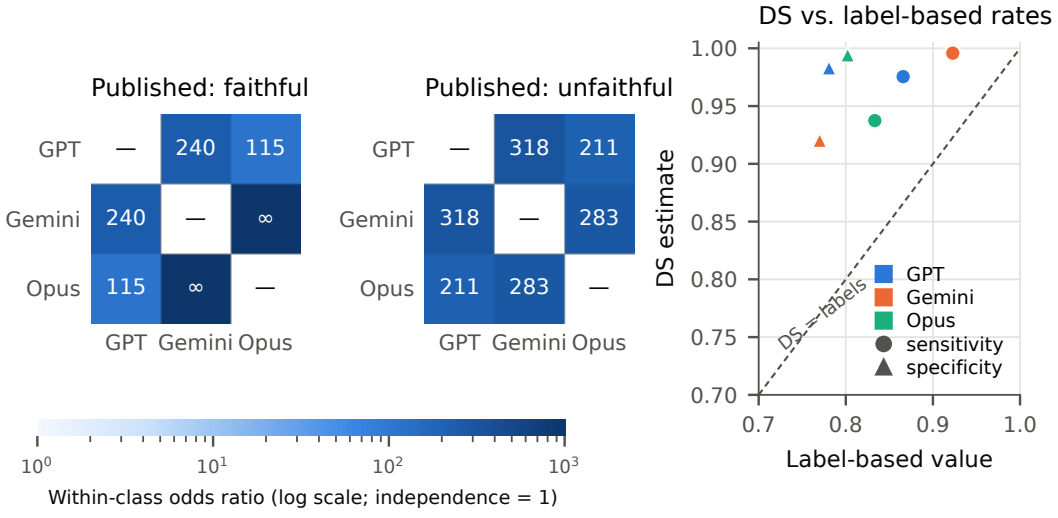


Figure 7: **The judges' votes are strongly related even among items with the same label** (CriticLeanBench). Left: odds ratio between two judges' votes within each label class. Independence within that group would give 1 (∞: an empty cell). Right: sensitivity and specificity as fitted by DS, against the values computed from the labels. Every fitted value exceeds its corresponding label-based value. Exact intervals: Table 9.

G Rankings for Inspection at Matched Budgets

This analysis is exploratory: the rankings were chosen after comparing several decision rules on the same benchmarks.

Definitions. Unselected items receive the three-judge majority label; on both benchmarks the DS decision ($P \geq 0.5$) equals the majority label. Rankings use judge outputs only.

- Model splits (2 score levels): split items first.
- DS posterior (8 and 5 levels): smallest $|p_i - 0.5|$ first, with posteriors fitted on all 450 CriticLeanBench and 100 MA-Align items and held fixed across budgets.
- Grade margin (12 and 11 levels): for a majority-accept item, $\min_c g_{ic} - 8.5$; for a majority-reject item, $8.5 - \max_c g_{ic}$, where c ranges over the nine calls; smallest first. The value 8.5 is the midpoint between grades 8 and 9; the acceptance cutoff remains 9.
- Nine-call vote share (secondary): smallest $|s_i - 0.5|$ first, where s_i is the share of calls with grade ≥ 9.

A budget of $b = \mathrm{round}(fN)$ items is reviewed. When the budget boundary falls inside a group of tied items, the remaining slots are filled uniformly at random; we report the expectation and the 2.5–97.5% range over tie-breaks.

Table 11: **Majority–label disagreements selected at equal review budgets** (expected counts under random tie-breaking). The majority has 71 disagreements on CriticLeanBench and 8 on MA-Align. 5% means 22 and 5 items, respectively. Tie-breaking ranges: Table 12.

	CriticLeanBench ($n = 433$)				**MA-Align** ($n = 96$)			
Ranking	5%	10%	20%	30%	5%	10%	20%	30%
Random	3.6	7.1	14.3	21.3	0.4	0.8	1.6	2.4
Model splits	10.7	20.3	26.0	31.6	0.8	1.6	3.0	3.6
DS posterior	13.9	20.2	25.2	30.1	3.0	3.0	3.0	4.8
Grade margin	11.8	21.2	30.9	40.0	1.0	1.0	2.9	5.2

Table 12: **Tie-breaking ranges** for Table 11. Entries are expected counts with the 2.5–97.5% range over tie-breaks in brackets (omitted when there is no tie at the boundary); item counts in parentheses.

	CriticLeanBench (71 disagreements)			**MA-Align** (8 disagreements)		
Ranking	10% (43)	20% (87)	30% (130)	10% (10)	20% (19)	30% (29)
Random (expected)	7.1	14.3	21.3	0.8	1.6	2.4
Model splits	20.3 [20–21]	26.0 [22–30]	31.6 [26–37]	1.6 [0–3]	3.0	3.6 [3–5]
DS posterior	20.2 [20–21]	25.2 [22–29]	30.1 [26–34]	3.0	3.0	4.8 [3–7]
Grade margin	21.2 [19–23]	30.9 [27–35]	40.0 [37–43]	1.0	2.9 [2–4]	5.2 [4–7]

Voting patterns. A binary DS fit assigns one posterior to each voting pattern (Table 13). The 41 split items contain 20 label disagreements and the 392 unanimous items contain 51. At a budget of 41, DS and model splits both select all split items and exactly 20 disagreements; beyond that, DS orders unanimous rejections before unanimous acceptances but cannot distinguish items within either group. Grade margin has more score levels and can separate some such items.

Secondary ranking. The nine-call vote share selects the following expected numbers of label disagreements at 10, 20 and 30% of items: 21.0 [20–22], 27.6 [24–32] and 33.0 [28–38] on CriticLeanBench; 2.0, 3.0 and 4.4 [3–6] on MA-Align.

Paired bootstrap. We resampled NL groups ($B = 1000$) with the DS posteriors held fixed and set budgets as fractions of each resample. Mean differences in label disagreements selected (95% percentile intervals, unadjusted):

Benchmark	Difference	10%	20%	33%
CriticLeanBench	Margin − DS	+1.3 (−1.4, 4.2)	+5.6 (1.6, 10.1)	+10.9 (2.8, 19.6)
	Margin − splits	+1.5 (−1.0, 4.7)	+4.9 (1.6, 8.7)	+9.3 (3.1, 15.6)
	DS − splits	+0.2 (−0.5, 2.4)	−0.7 (−2.5, 1.1)	−1.6 (−5.5, 2.5)
MA-Align	Margin − DS	−2.0 (−5.0, 0.0)	−0.5 (−3.4, 1.4)	+0.4 (−1.1, 2.0)
	Margin − splits	−0.5 (−2.2, 1.3)	0.0 (−2.3, 2.0)	+1.8 (0.0, 4.3)
	DS − splits	+1.5 (0.0, 3.9)	+0.4 (0.0, 1.6)	+1.5 (0.3, 3.2)

Confident items still contain disagreements. Among the 289 CriticLeanBench items whose nine grades are all 10 or all at most 7, 28 disagree with the published label, 14 of them acceptances.

On CriticLeanBench, the pattern DS ranks first, Gemini alone accepting, contains 11 label disagreements among 14 items, compared with 20 among all 41 split items (Table 13). DS assigns that pattern faithful probability 0.026 even though 11 of its items are labeled faithful. Its ordering can therefore be useful despite poorly calibrated probabilities. DS cannot distinguish items within a voting pattern, while grade margin retains differences among their nine grades.

Table 13: **CriticLeanBench voting patterns**, ordered by DS inspection priority (smallest $|p - 0.5|$ first). A: accepts; R: rejects. Faithful counts use published labels; disagreements compare the majority decision with those labels.

GPT/Gemini/Opus	Items	Labeled faithful	Disagreements	DS p
R / A / R	14	11	11	0.026
R / R / A	1	0	0	0.025
A / R / R	3	1	1	0.022
A / R / A	1	0	1	0.983
A / A / R	16	11	5	0.984
R / A / A	6	4	2	0.986
R / R / R	158	18	18	0.000010
A / A / A	234	201	33	0.999994

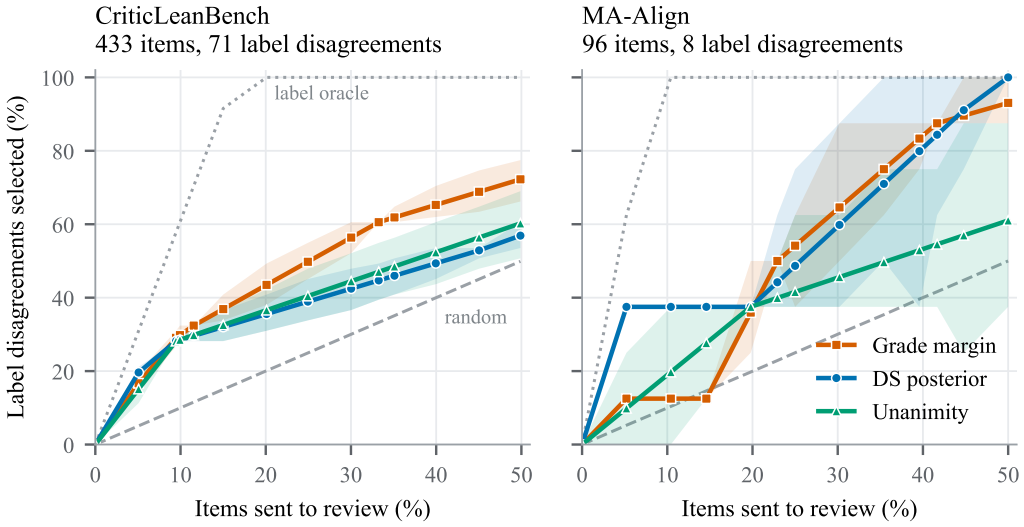


Figure 8: **Choosing what to inspect at equal workloads.** Vertical axis: fraction of all majority–label disagreements selected; horizontal axis: fraction of items reviewed. Lines average over random ties; bands show 2.5–97.5% tie-breaking ranges. Dashed: random selection; dotted: an ideal ordering using the labels.

H AI-ASSISTED AUDIT AND TARGETED CHECKS

We audited 75 CriticLeanBench items: 20 on which DS disagrees with the published label and 20 on which it agrees, drawn at random, and the remaining 35 of the 51 items on which all three judges contradict the label. Items were shown without labels or judge scores. Judgments were AI-assisted, so the audit is not an independent blind reference.

Across all 75 items there are 42 faithful, 24 unfaithful, and nine uncertain judgments. Among the 20 agreement items, all 17 definite judgments agree with both the label and the judges. Across the 51 unanimous disagreements, the judgments side with the judges on 44, with the label on two, and are uncertain on five. Table 14 splits the 35 items of the second set by direction.

Table 14: **Audit judgments on 35 unanimous disagreements**, by direction. The remaining 16 unanimous disagreements (four labeled faithful, twelve labeled unfaithful) belong to the first set of 40.

Published label / judges	Items	Faithful	Unfaithful	Uncertain
Unfaithful / all accept	21	18	1	2
Faithful / all reject	14	0	12	2

Checks on individual items. CR19 has a positive label, but the source's positive indices become unrestricted naturals in the conclusion; a Lean-verified counterexample is given in Section K. CR06 (`test/000259`) asks about triangle angles satisfying $\sin^2 A + \sin^2 B + \sin^2 C = 2$, but the formal statement omits $A + B + C = \pi$. Setting all three angles to $\arcsin\sqrt{2/3}$ satisfies the formal constraints and gives $\cos A + \cos B + 2\cos C = 4/\sqrt{3} > \sqrt{5}$, exceeding the claimed maximum; this supports the negative label, and our first judgment of this item (faithful) was wrong. CR15 is accepted under a rounding convention, which does not make its exact equality true. The judgments of R05 (a French source), R24 (an added answer-formatting task), R26 (a rounded answer stated as an exact value) and R31 (a flawed proof passage alongside a stated theorem) remain uncertain. The checks for R15, R14, R31, R24 and R26 appear in Section K. R24's original source is the 2000 AMC 12, Problem 15, which asks only for the sum;[1] R26's nearest-integer instruction appears in the official 2017 HMMT Guts solutions, Problem 9.[2]

H.1 USING THE AUDIT JUDGMENTS AS THE REFERENCE

We replace the published label by the audit judgment on the 66 items with a definite judgment, keep the published label elsewhere, and omit the nine uncertain items. DS posteriors and rankings do not change, because labels do not enter DS fitting. Of the 433 items, 424 have a resulting reference, which differs from the published label on 46 of them.

Table 15: **Published versus audit-based reference**, on the same 424 items. Accuracy is agreement with the reference; lower Brier is better.

	Accuracy (%)		Brier	
Method	Published	Audit-based	Published	Audit-based
GPT	84.0	94.8	.152	.047
Gemini	86.8	97.6	.132	.025
Opus	82.8	92.7	.168	.067
Majority	84.7	95.5	.134	.029
DS	84.7	95.5	.151	.043

[1] `https://f.hubspotusercontent30.net/hubfs/981338/Math%20Competition%20Problems_PDFS/2000AMC12.pdf`.

[2] `https://s3.amazonaws.com/hmmt-archive/february/2017/February2017GutsSolutions.pdf`.

The rise in agreement is largely by construction, since the audit targeted disagreements and mostly sided with the judges. The comparison between methods is unchanged: DS and majority voting have identical verdicts, and DS still has the higher Brier score. Among the 253 items with posterior above 0.8, the share with a faithful reference rises from 85.0% to 97.6%, so conclusions about calibration depend on the reference.

H.2 What each evidence source establishes

Table 16: **Different signals support different conclusions.**

Signal	Supports	Does not establish alone
Human or LLM agreement	Consensus under the supplied instructions	Correctness
DS posterior	Inference under the model's assumptions, including conditional independence of votes	Validated confidence in an item's faithfulness
Published-label agreement	Consistency with a benchmark reference	An error-free semantic reference
AI-assisted audit	Specific issues for further checking	Independent validation of the labels
Claim-level Lean check	The formal claim actually checked	Faithfulness of every part of a translation

I LLM-Only Rankings and Human Disagreement

This analysis is exploratory and reuses the existing model and human judgments.

Fitting and ranking. Each model casts a binary vote (grade ≥ 9) by majority of three valid calls; an incomplete triple gives a missing vote. We fit LLM-only DS separately on all 116 A and 92 B items using the available votes; no human grade, note or abstention enters fitting or ranking. The fitted faithful shares are 0.18 on A and 0.76 on B. Rankings are as on the benchmarks (Section G). Items with all three model votes form the ranking pool (109 on A, 90 on B); the comparison further requires at least two human grades, leaving 107 A and 89 B items. Human disagreement means at least one grade ≥ 9 and another < 9; R3's abstentions are analyzed separately. Budgets are 10%, 20% and 30% of the eligible pool, and ties at the boundary are broken at random. Intervals use 2,000 paired resamples of NL groups within each panel, with rankings held fixed; they are not adjusted for multiple comparisons. AUC counts ties as one half.

Table 17: **Intervals for Table 4** (20% budget). Expected percentages under random tie-breaking, with 95% paired-bootstrap intervals. Random selection has the pool's baseline rate.

Ranking	**Panel A** (21 of 107)	**Panel B** (18 of 89)
Random / pool baseline	43.9%	28.1%
Model splits first	73.7% [60.5, 87.1]	48.1% [29.2, 66.7]
DS uncertainty	71.8% [56.1, 87.2]	44.9% [21.8, 66.7]
Grade margin	77.2% [62.0, 90.5]	53.7% [30.1, 73.9]

Table 18: **Expected human-disagreement rates among selected items.** Each entry is a percentage; the random baseline equals the eligible pool rate (43.9% A, 28.1% B).

	Panel A ($n = 107$)			**Panel B** ($n = 89$)		
Ranking	10% (11)	20% (21)	30% (32)	10% (9)	20% (18)	30% (27)
Random	43.9	43.9	43.9	28.1	28.1	28.1
Model splits	73.7	73.7	73.7	48.1	48.1	48.1
DS uncertainty	72.7	71.8	73.2	30.0	44.9	48.1
Grade margin	83.0	77.2	75.0	52.1	53.7	46.9

Table 19: **AUC for predicting observed human disagreement**, with descriptive 95% fixed-ranking bootstrap intervals. Random ordering has AUC 0.5.

Ranking	Panel A	Panel B
Model splits	0.715 [0.629, 0.795]	0.651 [0.536, 0.769]
DS uncertainty	0.723 [0.635, 0.808]	0.642 [0.522, 0.769]
Grade margin	0.807 [0.728, 0.879]	0.688 [0.559, 0.815]

At 20%, DS exceeds the random baseline by 27.9 percentage points on A (paired interval 13.0–42.7) and 16.9 on B (-4.1 to 35.9). The DS-minus-split intervals include zero on both panels.

Sensitivity analyses. Restricting to items with all three human ratings leaves 85 A and 68 B items, with baseline disagreement rates of 43.5% and 26.5%; at 20%, DS selects 68.9% and 45.2%, versus 70.0% and 47.4% for model splits. Removing R1's style deductions on B reduces the disagreements from 25 to 17 among the same 89 items; DS then selects 35.4% at 20%, versus a 19.1% baseline and 33.3% for model splits (DS minus splits: -12.8 to 11.4 points). For R3's abstentions (24/109 on A, 21/90 on B), DS AUC is 0.496 [0.372, 0.623] on A and 0.693 [0.575, 0.799] on B.

B081. The grades are R1/R2/R3 $= 5/10/10$; GPT $= 7/7/7$, Gemini $= 10/10/10$, and Opus $= 5/5/4$. LLM-only DS gives faithful probability 0.066. In the 89-item B pool it is selected with probability 0 at 10%, $8/11$ at 20%, and 1 at 30%; grade margin includes it at 20%.

J Ranking Candidates for Reconsideration

This exploratory analysis asks whether LLM judgments help locate earlier screening rejections supported by the existing human majority. It reuses the LLM-only DS fits from Section I; no parameters are refitted and no new judgments are collected.

Evaluation pool and outcome. We require all three model votes and a decisive human majority: at least two original grades ≥ 9 for acceptance or at least two below 9 for rejection. Two opposing grades with the third missing are undecided. Of the 109 Panel-A items with complete model votes, 97 have a decisive human majority, which accepts 18 and rejects 79. Panel B has 82 eligible items, with 67 accepted and 15 rejected. Human ratings determine evaluation eligibility and outcomes, but do not enter the ranking scores. The outcome is support from the existing human majority, rather than independently verified faithfulness.

Rankings and budgets. We rank by decreasing DS faithfulness probability, the fraction of models voting accept, or the mean of all nine model grades. This differs from ranking by uncertainty to find reviewer disagreements. DS posteriors are rounded to 12 decimal places to merge numerical ties. Each ranking selects 10%, 20%, or 30% of the eligible pool, rounded to the nearest item. Counts are exact expectations under uniform random selection among items tied at the cutoff. The random baseline is the review count multiplied by the pool's human-majority acceptance rate. We report both the human-majority acceptance rate within the selected list and the fraction of all human-majority acceptances selected.

Uncertainty. Intervals use 2,000 paired bootstrap resamples of exact-NL groups within each eligible pool (95 groups on A, 79 on B). Each replicate samples the original number of groups with replacement, keeps their eligible items together, and recomputes the review count for the resulting pool size. All methods use the same resample, with ranking scores held fixed. The 95% percentile intervals describe sampling variation conditional on the fitted scores and observed human outcomes; they do not include DS fitting uncertainty or adjustment for multiple comparisons. The analysis records seeds 20260925 (A) and 20260926 (B), input hashes, and eligible item IDs.

At 20%, DS selects 11 of the 18 human-majority acceptances: 61.1%, with an interval of [41.7, 79.9]. The gain in selected acceptance rate over random selection is 39.3 percentage points, with a paired interval of [16.7, 55.4]. The differences from vote share and mean grade are both zero at

Table 20: **Human-majority acceptance among selected Panel-A items at a 20% review budget.** Each method selects 19 of 97 eligible earlier screening rejections. Intervals use paired NL-group bootstrap resampling.

Ranking	Selected acceptance rate	95% interval
Random	18.6%	[10.4, 27.1]
DS probability, highest first	57.9%	[29.2, 79.5]
Model acceptance vote share	57.9%	[31.1, 79.6]
Mean grade	57.9%	[26.3, 80.0]

this budget, with intervals of $[-5.1, 0.0]$ and $[-7.6, 8.0]$ percentage points, respectively. Across the three budgets, the methods have similar yields (Table 22).

Panel B comparison. Panel B includes both earlier acceptances and rejections, so ranking its full eligible pool addresses finding human-supported candidates in a mixed cohort. Human-majority acceptance is already 81.7% in this pool. DS and vote share select lists with 98.1% human-majority acceptance at all three budgets; mean grade selects lists with 100% in these data (Table 21). At 20%, the DS selected acceptance rate has interval [93.6, 100.0], and its gain over random selection is 16.4 percentage points [7.6, 25.7]. The three methods' point estimates again give no advantage to DS.

Table 21: **Human-majority acceptances selected on Panel B.** Expected counts among 82 eligible items, of which 67 are accepted by the human majority.

Ranking	10% (8 items)	20% (16 items)	30% (25 items)
Random	6.5	13.1	20.4
DS probability, highest first	7.8	15.7	24.5
Model acceptance vote share	7.8	15.7	24.5
Mean grade	8.0	16.0	25.0

Table 22: **Finding human-supported candidates among earlier screening rejections.** Expected number of human-majority acceptances selected from Panel A (18 among 97 eligible items). Each ranking uses only LLM outputs; ties are averaged.

Ranking	10% (10 items)	20% (19 items)	30% (29 items)
Random	1.9	3.5	5.4
DS probability, highest first	7.3	11.0	13.3
Model acceptance vote share	7.3	11.0	13.6
Mean grade	7.3	11.0	13.0

K Case Studies: From Disagreement to Specific Evidence

These ten illustrative cases connect the aggregate results to concrete source statements, Lean code, and judgments; they are examples, not a frequency sample. Grades are shown in call order, and a model accepts by majority at grade ≥ 9. Panel cases have no trusted ground truth. Lean checks use Lean 4.28.0 and contain no `sorry`.

CR19: A positive label with an invalid zero-index conclusion

CriticLeanBench; `test/000127`.

Source statement

Let $f(m,n)$ satisfy $f(1,n) = f(m,1) = 1$ $(m,n \in \mathbf{N}^*)$, and when $m,n \geqslant 2$, we have

$$f(m,n) \leqslant f(m,n-1) + f(m-1,n).$$

Prove that $f(m,n) \leqslant \mathrm{C}_{m+n-2}^{m-1}$.

Lean formalization

```
import Mathlib

open Finset

theorem algebra_539172 {f : ℕ → ℕ → ℕ}
    (h₁ : ∀ n, f 1 n = 1) (h₂ : ∀ m, f m 1 = 1)
    (h₃ : ∀ m n, 2 ≤ m → 2 ≤ n →
      f m n ≤ f m (n - 1) + f (m - 1) n) :
    ∀ m n, f m n ≤ Nat.choose (m + n - 2) (m - 1) := by sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 6/6/6; Gemini: 7/7/7; Opus: 4/4/4.

Closer look

The source restricts m,n to positive integers. The formal conclusion quantifies over all natural numbers, including zero. Set $f(0,0) = 2$ and $f(m,n) = 1$ elsewhere. The two boundary hypotheses hold, and the recurrence assumption for $m,n \geq 2$ reduces to $1 \leq 1+1$. At $(0,0)$, however, the conclusion requires $2 \leq \binom{0}{0} = 1$, since natural-number subtraction truncates at zero.

Evidence and scope

A Lean proof negates the universal closure of the displayed statement, preserving its hypotheses and conclusion. It compiles without `sorryAx`. Together with the positive-index restriction in the source, this establishes a translation mismatch in an item labeled faithful.

R15: Orthogonality becomes an impossible hypothesis

CriticLeanBench; `test/000266`.

Source statement

Prove that if $\overrightarrow{a} = (x, 2)$ and $\overrightarrow{b} = (x - 1, 1)$ and $(\overrightarrow{a} + \overrightarrow{b}) \perp (\overrightarrow{a} - \overrightarrow{b})$, then $x = -1$.

Lean formalization

```
import Mathlib


theorem prove {x : ℝ} (a b : ℝ × ℝ)
 (ha : a = (x, 2)) (hb : b = (x - 1, 1))
 (h : (a.1 + b.1, a.2 + b.2) • (a.1 - b.1, a.2 - b.2) = 0) :
 x = -1 := by sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 4/4/4; Gemini: 7/5/4; Opus: 3/3/3.

Closer look

The source's orthogonality condition is satisfiable: at $x = -1$, the dot product $(a + b) \cdot (a - b)$ is zero. The formal hypothesis instead uses the scalar action on a product. After substituting the vectors, its second component requires $(2 + 1)(2 - 1) = 0$, hence $3 = 0$. The hypothesis therefore excludes every instance of the intended problem.

Evidence and scope

Lean proves that the displayed hypothesis is impossible after the substitutions. The resulting implication is vacuously true; we are not giving a counterexample to that implication. The source comparison identifies the error: componentwise multiplication does not represent the stated orthogonality condition.

R14: Integer inputs change the meaning of division

CriticLeanBench; `test/000490`.

Source statement

Demonstrate that there are exactly **224** integers N satisfying the following two conditions:

1. **Existence of a real solution:** There exists a real number x such that:

$$x^2 + \frac{11}{10}x + \frac{N}{1110} < 0.$$

2. **Positivity for all integers:** For every integer n, the following inequality holds:

$$n^2 + \frac{11}{10}n + \frac{N}{1110} > 0.$$

Prove that the number of such integers N is **224**.

Lean formalization

```
import Mathlib

theorem Count :
 Set.ncard {N : ℤ | ∃ x : ℝ, x ^ 2 + (11 / 10) * x + N / 1110 < 0 ∧
 ∀ n : ℤ, n ^ 2 + (11 / 10) * n + N / 1110 > 0} = 224 := by sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 4/5/5; Gemini: 5/6/5; Opus: 3/3/4.

Closer look

The source uses the same real-valued quadratic expression in both conditions, evaluating it first at a real x and then at integer n. In the Lean condition quantified over $n : \mathbb{Z}$, the arithmetic is instead integer arithmetic. In particular, $11/10 = 1$, and $N/1110$ is integer division. Thus the second condition is not obtained merely by restricting the argument of the first real-valued expression.

Evidence and scope

A Lean equality check identifies the second condition with its explicit integer-division form. It verifies the elaborated arithmetic, not the final claim that the set has cardinality 224.

R31: A proof-like source and an integer-division expression

CriticLeanBench; `test/000139`.

Source statement (excerpt)

Proof: For any real number x, we need to prove that $\sum_{k=1}^{n} \frac{[kx]}{k} \leqslant [nx]$, where $[x]$ represents the greatest integer not exceeding x.

Lean formalization

```
import Mathlib

open Real Int Nat Set
open scoped BigOperators

theorem algebra_58859 (x : ℝ) (n : ℕ) :
    ∑ k in Finset.Icc 1 n, ⌊k * x⌋ / k ≤ ⌊n * x⌋ := sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 5/5/5; Gemini: 7/6/6; Opus: 4/4/4.

Closer look

The released source continues with an attempted induction proof; the displayed excerpt is its initial statement. One later step asserts $\lfloor mx + x \rfloor \leq \lfloor mx \rfloor + \lfloor x \rfloor$, which fails for $m = 1$, $x = 3/5$. Separately, Lean interprets $\lfloor kx \rfloor / k$ as integer division, rather than the real-valued quotient in the written sum. A flawed proof passage and a changed formal expression are distinct issues.

Evidence and scope

A Lean check establishes the integer-division type. The check does not adjudicate every interpretation of the source theorem. Our judgment of this item is uncertain.

CR15: Rounded reporting versus an exact root

CriticLeanBench; `test/000203`.

Source statement

Find the smallest positive solution of the equation $\sum_{n=0}^{\infty} \frac{(-x)^n}{(n!)^2} = 0$ and provide the answer with four digits of the fractional part. Prove that the answer is 1.4458.

Lean formalization

```
import Mathlib

open Real

theorem solution :
    IsLeast {x : ℝ | 0 < x ∧
      ∑' n : ℕ, (-x) ^ n / (Nat.factorial n) ^ 2 = 0}
      (1.4458) := by sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 8/7/7; Gemini: 7/5/7; Opus: 4/4/3.

Closer look

The source asks for four digits after the decimal point. The formalization instead asserts that 1.4458 itself is the least positive root. Numerical evaluation gives $S(1.4458) \approx -1.51514 \times 10^{-6}$, while the smallest positive root is approximately $1.445796\ldots$ and rounds to 1.4458. Thus reporting precision and exact equality are different requirements.

Evidence and scope

This is a numerical check, not a Lean-verified proof. Our audit accepts the answer under a permissive rounding convention, which does not make the exact equality true. The case illustrates a grading convention that should be specified explicitly, rather than an unqualified claim that either the label or the judges must be wrong.

R26: A finite iteration with a rounded answer

CriticLeanBench; `test/000372`.

Source statement

Jeffrey writes the numbers 1 and $100000000 = 10^8$ on the blackboard. Every minute, if x, y are on the board, Jeffrey replaces them with $\frac{x+y}{2}$ and $2\left(\frac{1}{x}+\frac{1}{y}\right)^{-1}$. After 2017 minutes the two numbers are a and b. Find $\min(a, b)$ to the nearest integer. Prove that the answer is: 10000

Lean formalization

```
import Mathlib

open scoped BigOperators

theorem omni_theorem_435 (x y : ℕ → ℝ)
    (hx0 : x 0 = 1) (hy0 : y 0 = 10 ^ 8)
    (hxn : ∀ n, x (n + 1) = (x n + y n) / 2)
    (hy : ∀ n, y (n + 1) = 2 * ((1 / x n) + (1 / y n)) ^ (-1 : ℤ)) :
    IsLeast {z | ∃ a b, z = min a b ∧ a = x 2017 ∧ b = y 2017} 10000
  := by sorry
```

Judgments

Published label: faithful.

LLM grades: GPT: 6/5/6; Gemini: 5/5/5; Opus: 3/3/4.

Closer look

The two updates are the arithmetic and harmonic means. For positive unequal a, b, they preserve the product ab and differ by $(a-b)^2/[2(a+b)] > 0$. Starting from 1 and 10^8, the values therefore remain unequal with product 10^8 at every finite step. Their minimum is strictly below 10^4, including at step 2017. The formal conclusion asserts exact equality to 10000, whereas the source requests the nearest integer.

Evidence and scope

The finite-step argument is mathematical, not Lean-verified. Official 2017 HMMT Guts solutions (Problem 9) retain the nearest-integer requirement. Like CR15, this case contrasts rounding with exact equality. Our judgment of this item is uncertain.

R24: The released task differs from the original question

CriticLeanBench; `test/000175`.

Source statement

Let f be a function for which $f\left(\frac{x}{3}\right) = x^2 + x + 1$. Find the sum of all values of z for which $f(3z) = 7$. The original answer is in the form of $-\frac{k}{m}$, where k and m are integers. Please give the value of $k + m$. prove that answer is 10

Lean formalization

```
import Mathlib

theorem sumOfZs (f : ℝ → ℝ)
    (h : ∀ x, f (x / 3) = x ^ 2 + x + 1) :
    ∃ k m : ℤ,
      (∑ᶠ z ∈ {x | f (3 * x) = 7}, z = -k / m) ∧
      k + m = 10 := by sorry
```

Judgments

Published label: unfaithful.

LLM grades: GPT: 10/10/10; Gemini: 10/9/10; Opus: 9/9/9.

Closer look

The original 2000 AMC 12 Problem 15 asks only for the sum of solutions; the answer is $-1/9$. The released text adds the representation $-k/m$ and the request for $k + m$. Substitution gives $81z^2 + 9z - 6 = 0$, so the root sum is indeed $-1/9$. A unique numerator-plus-denominator answer needs conventions such as lowest terms and a positive denominator, not merely that k, m are integers.

Evidence and scope

Comparison with the original contest paper confirms the added task. The Lean statement existentially chooses k, m, so the normalization concern alone is not a complete proof that this NL–Lean pair is unfaithful. Our judgment is uncertain; the case shows ambiguity introduced in task construction rather than a verified label error.

B081: Two reviewers miss a strengthened hypothesis

Panel B; B081; no trusted ground-truth label.

Source statement

Suppose $\sum a_n$ converges absolutely and $\sum b_n$ converges. Show that $c_n = \sum_{k=0}^{n} a_k b_{n-k}$ defines a convergent series with sum equal to the product of sums.

Lean formalization

```
import Mathlib

theorem jirilebl_ra_ch_seq_ser_4256
    (a b : ℕ → ℝ) (s t : ℝ)
    (ha : HasSum a s)
    (ha_abs : Summable (fun n => ‖a n‖))
    (hb : HasSum b t) :
    HasSum (fun n => ∑ k ∈ Finset.range (n + 1), a k * b (n - k)) (s
  * t) := by sorry
```

Judgments

Human grades: R1: 5, R2: 10, R3: 10.

LLM grades: GPT: 7/7/7; Gemini: 10/10/10; Opus: 5/5/4.

Closer look

The source requires absolute convergence for $\sum a_n$, but only ordinary convergence for $\sum b_n$. In Mathlib, `HasSum b t` over $\mathbb{R}$ concerns an unordered sum and implies absolute summability. Thus the formal hypothesis on b is stronger: it excludes conditionally convergent series covered by the source.

Evidence and scope

Lean verifies the implication from `HasSum b t` to summability of the absolute values. This establishes the strengthened hypothesis, not complete faithfulness of the translation. R2 and R3 accepted the item despite this issue.

B099: A judge misreads the scope of a sum

Panel B; B099; no trusted ground-truth label.

Source statement

Let $z \in \mathbb{C}$. Show that for all $n \in \mathbb{N}$, $\left|\sum_{k=0}^{n} \frac{z^k}{k!} - e^z\right| \leq \frac{|z|^{n+1}}{(n+1)!} e^{|z|}$.

Lean formalization

```
import Mathlib

theorem jirilebl_ca_ca_19562 (z : ℂ) (n : ℕ) :
    ‖∑ k ∈ Finset.range (n + 1), z ^ k / (k.factorial : ℂ) - Complex.
  exp z‖ ≤
    ‖z‖ ^ (n + 1) / ((n + 1).factorial : ℝ) * Real.exp ‖z‖ := by sorry
```

Judgments

Human grades: R1: 10, R2: 10, R3: 10.

LLM grades: GPT: 10/5/5; Gemini: 10/10/10; Opus: 10/10/10.

Closer look

Two GPT calls objected that `Complex.exp z` was subtracted inside every summand. In the notation actually used here, the subtraction is outside the finite sum: the norm encloses the partial sum minus e^z, as in the source. The repeated objection therefore relies on an incorrect parse.

Evidence and scope

A Lean equality check confirms the scope of the sum. It refutes this particular GPT objection, not every possible objection to the formalization. Repeating the same explanation twice does not make it correct, and a majority over the three GPT calls preserves the error in this case.

B026: Repairing a false source changes the translation

Panel B; B026; no trusted ground-truth label.

Source statement

Let $z \in \mathbb{C}$. Prove that $|\sinh z| \leq \frac{1}{2}e^{|z|}$ and $|\cosh z| \leq \frac{1}{2}e^{|z|}$.

Lean formalization

```
import Mathlib

open Complex

lemma norm_sinh_le (z : ℂ) : ‖sinh z‖ ≤ Real.exp ‖z‖ / 2 := by sorry

lemma norm_cosh_le (z : ℂ) : ‖cosh z‖ ≤ Real.exp ‖z‖ := by sorry

theorem jirilebl_ca_ca_15733 (z : ℂ) :
    ‖sinh z‖ ≤ Real.exp ‖z‖ / 2 ∧ ‖cosh z‖ ≤ Real.exp ‖z‖ := by sorry
```

Judgments

Human grades: R1: 8, R2: 8, R3: abstained.

LLM grades: GPT: 7/7/7; Gemini: 7/7/7; Opus: 6/6/6.

Closer look

The source's cosh bound is false at $z = 0$: it requires $1 \leq \frac{1}{2}$. The formalization removes the factor $1/2$ from that bound, both in its helper lemma and in the final conjunction. This avoids the displayed counterexample to the source, but also changes what was asked. Mathematical repair and literal statement translation are different tasks.

Evidence and scope

A Lean check confirms that the source bound fails at zero; the displayed texts show the removed factor. This does not certify the rest of either statement. R1's score is 8, or 10 after removing style deductions, so removing a style deduction does not establish semantic correctness.