BEYOND SELECTED BREAKTHROUGHS: EVALUATING LANGUAGE MODELS ON RANDOMLY SAMPLED MATHEMATICAL CONJECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

The last year has seen a shift in AI-assisted mathematics, with frontier language models producing proofs and counterexamples for long-standing open problems. However, curated benchmarks and selected breakthroughs provide an incomplete picture of LLMs performance and the computational resources and problem-specific mathematical guidance needed to achieve reported results. To address this, we introduce a new dataset, OpenConjecture, which contains most conjectures from math papers published on the arXiv since February 2026. We apply frontier LLMs to a sample of 200 of these using a reproducible protocol without problem-specific mathematical guidance. Taken together, the various models produce candidate solutions for roughly 60% of the conjectures, with counterexamples more common than proofs. We then perform a qualitative analysis of solution characteristics such as novelty, and connections to existing literature. Our analysis distinguishes proposed conjecture resolutions that expose missing assumptions or recover existing results from those that may contribute new mathematics. Finally, we document the prompts and computational resources required, and release the model responses to support further study.

1 INTRODUCTION

The last year has seen a dramatic shift in the capabilities of frontier AI systems in research-level mathematics. The disproof of the Erdős unit distance conjecture (OpenAI, 2026), Jacobian conjecture (Sircar, 2026), the recent Navier–Stokes blowup result (OpenAI, 2026), and other announcements of AI-assisted advancements (OpenAI, 2026) have demonstrated that AI systems are able to produce proofs, as well as counterexamples, to difficult mathematical problems. This has become clear not only from announcements from frontier labs but also through carefully curated benchmarks such as First Proof (Abouzaid et al., 2026a;b) and Benchmarks in Leipzig (Balakin et al., 2026). These and similar benchmarks have helped assess model performance on collections of problems whose solutions are known but not publicly available online, and often provide detailed information about prompting and cost.

Still, the rapid progress has left many questions unanswered. Do LLMs more often disprove conjectures by finding counterexamples, or prove them? How many conjectures can be resolved with a single naïve prompt, and how often is repeated prompting or domain expertise needed? Are LLM solutions primarily driven by broader awareness of the literature such that little novelty is typically introduced? We are interested in how these and related questions would be answered on a *random* sample of conjectures.

This motivates OpenConjecture, a dataset of conjectures from recent arXiv mathematics preprints. In this paper we introduce this dataset and use it as a source of conjectures drawn 'from the wild'. We study 200 conjectures, one drawn from each of 200 randomly sampled papers. We then prompt several versions of GPT models, specifically GPT-5.6-Sol, GPT-5.6-Sol Pro, and GPT-6-Astra Pro, to produce proofs or counterexamples across three passes, each of which recieves the previous pass's unresolved attempts as context. We analyze the resulting output to better understand the qualities of LLM mathematical ability beyond correctness. As such our purpose is not to announce a long list of new results, but rather see what we can learn about the types of solutions these systems produce.

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.

Out of courtesy to the reader, we keep the LLM-generated mathematics to a minimum, including examples only as needed to illustrate our main points.

We find that the models' proposed conjecture resolutions vary significantly in what they contribute mathematically. For example, of the proposed counterexamples from the first and second passes that passed an LLM verification review, roughly 60% were classified as "off-the-shelf" and the remainder as "standard constructions". Some of the counterexamples appeared to take advantage of missing hypotheses in the conjecture statement rather than offer a mathematically interesting construction. In contrast, the counterexamples from the third pass typically involved nontrivial modifications or more involved arguments that the proposed object does in fact refutee the conjecture. Findings like these illustrate that technical correctness is just one dimension of LLM mathematical output.

Concretely, this paper makes the following contributions:

- We introduce the OpenConjecture dataset, which gathers conjectures from recent arXiv preprints and makes them publicly available for analysis.
- We prompt frontier GPT models, in three passes, to solve a sample of 200 conjectures drawn from this dataset. We find that models propose counterexamples to 70 conjectures and proofs to 55 conjectures, and that 84% of these proposed resolutions hold up to LLM review.
- We analyze the mathematical content of proposed solutions, including the novelty of counterexample constructions and the relationship of proposed proofs to the existing literature.

While we do not attempt expert verification or formalization of the conjecture attempts or use this as a benchmark to compare models, we encourage extensions of this work: OpenConjecture is available on HuggingFace[1] and model responses to the 200 sampled conjectures will be made available separately.

2 Related Work

The field of AI for math has grown significantly in recent years. We refer the reader to (Ju & Dong, 2026) for a survey and (Lee, 2026) for a curated list of papers in the area. Here, we focus on the related work that most directly informs our methodology.

Similar to OpenConjecture, **LemmaBench** (Peyronnet et al., 2026) and **arXivMath** (Dekoninck et al., 2026) also source problems from recent arXiv papers. Both cite the high risk of data contamination as a motivator. LemmaBench rewrites lemmas from recent arXiv preprints into self-contained statements. ArXivMath was generated by asking an LLM to create questions from arXiv abstracts, rejecting abstracts that do not give rise to a self-contained problem. Then they have several additional filters and preprocessing steps: they use an LLM to add missing conditions, and to check (from the full paper in markdown) whether the answer is inferrable from prior work. Like these datasets, OpenConjecture and our corresponding analysis aims to capture mathematics 'in the wild'. Unlike these datasets however, OpenConjecture aims to collection conjectures made by human mathematicians, rather than lemmas or questions synthetically generated via an abstract.

The **First Proof** batches each consist of 10 research-level questions. The questions span a variety of fields in mathematics and are problems that arose in the research process of the contributors but whose proofs had not been published. The authors describe the first batch of 10 problems (Abouzaid et al., 2026a) as an informal experiment, and the second batch of 10 problems as a more formal benchmark. In the second batch, they tested and graded four systems: Chat GPT-5.5 Pro and three harnesses. Solutions were refereed and graded by experts. Some particularly noteworthy findings included: (i) at least seven of the 10 problems received at least one passing grade; (ii) in three problems, the argument was not the same as the known solution; (iii) in several solutions, the failures to properly attribute were significant enough that if a human submitted the solutions it would be considered plagiarism.

Other notable research-level benchmarks include Benchmarks in Leipzig (Balakin et al., 2026), which consists of 100 research-level questions with unique, well-defined answers, contributed from

[1] https://anonymous-hf.com/a/gc1rpn4bihkw/.

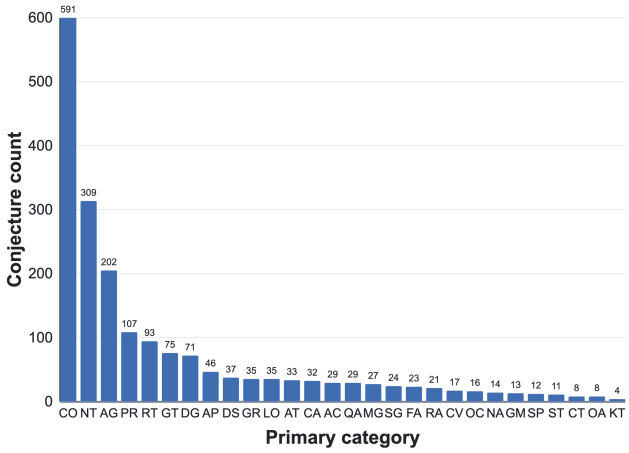

Figure 1: All open conjectures with a math primary subject classification in the June 21st snapshot of OpenConjecture. Categories with a single conjecture are omitted.

49 mathematicians, IMProofBench (Schmitt et al., 2025), Soohak (Son et al., 2026), Riemann-Bench (Garre et al., 2026), and FrontierMath (Glazer et al., 2024).

3 OpenConjecture

The OpenConjecture dataset is created by ingesting recent arXiv math preprints and extracting conjecture blocks. We use GPT-5 Mini to label if the conjecture is actually being posed by the authors, as opposed to a conjecture that is being stated because the paper has resolved it or given as background motivation. We also use GPT-5 Mini to score conjectures for their mathematical interestingness and near-term solution viability statistics which may be of independent interest based on what they say about the landscape of actual conjectures. The first OpenConjecture snapshot was published March 17, 2026, with most papers uploaded after February 2nd, 2026. It is updated weekly.

For our case study, we use the June 21st snapshot, which contains 3030 conjectures (from 1675 papers) that were labeled as open by GPT-5 Mini. We filtered an additional 769 conjectures that were from papers that did not have a math primary subject classification. These papers were cross-listings or updates rather than papers from the intended window.

Motivated by the observation from arXivMath (Dekoninck et al., 2026) that even after their initial filtering, a PhD mathematician review of 20 questions found that 30% of the questions were potentially answerable using prior work cited, we re-ran the open/not-open filtering with GPT-5.4 on the 2261 remaining conjectures. In this step, 337 conjectures were reclassified as not real, leaving us with 1924 conjectures labeled as open.

The majority of remaining 1924 conjectures come from combinatorics, number theory, and algebraic geometry, as shown in Figure 1. The overall population of papers by subfield in the entire snapshot is shown in Figure 6 the Appendix. While combinatorics is the most prevalent, there are four categories with more papers than number theory and algebraic geometry (analysis of PDEs, optimization and control, numerical analysis, and probability). Other descriptive statistics related to the dataset and further preprocessing details can be found in Appendix A.

4 Conjecture Solution Pipeline

In this section we describe our methodology for further preprocessing and sampling conjectures, and how the models were prompted to solve the conjectures.

Sampling We sample 200 papers uniformly from papers containing eligible conjectures, and then sample one conjecture from each paper. The distribution of conjectures by primary category is

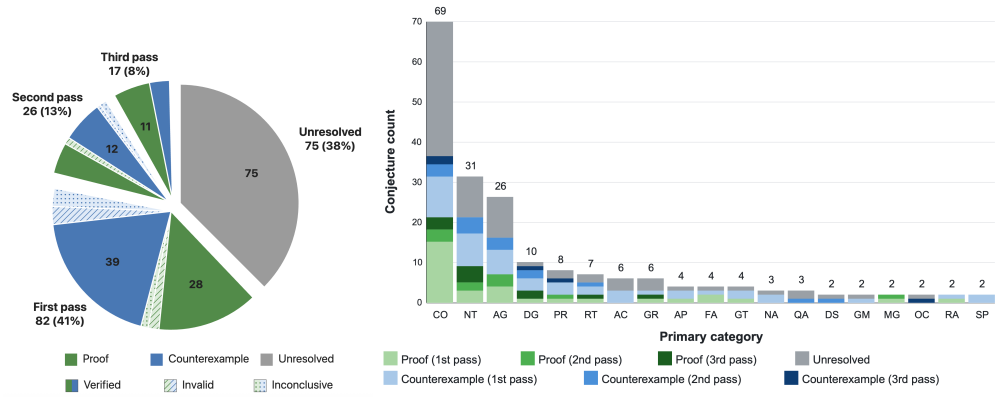

Figure 2: Left: Final status of conjectures and the pass in which a resolution was proposed; hatching indicates the LLM verification outcome. Right: Final status by math subject classification, with verification status omitted. The seven single-conjecture categories are not shown (AT, CT, CV, HO, LO, OA, and SG, with outcomes: three counterexamples, two proofs, and two unresolved).

shown in Figure 1. We chose this approach, rather than sampling a fixed number of conjectures by primary category because we wanted our results to reflect the global population of math conjectures in arXiv preprints without overweighting subfields that generate few conjectures.

Solver Our solution procedure involves passing each conjecture to three LLMs of increasing capability: the first pass uses GPT-5.6-Sol, the second pass uses GPT-5.6-Sol Pro, and the third pass uses GPT-6-Astra Pro. All runs used reasoning effort `xhigh`. This design was intended to approximate a workflow of a mathematician using frontier models under cost constraints: starting with a strong but not the most expensive model, and increasing the strength of the model in subsequent attempts if the conjecture is not resolved.

In the first pass we provided the conjecture, the full paper, and the instruction to "make a sustained attempt to prove or refute" the conjecture to GPT-5.6-Sol. The model could return one of three verdicts: `proof`, `counterexample`, or `unresolved`. It was also instructed to cite any external results used. The complete prompt is given in Appendix C.1. We kept the prompt relatively simple, following the example of FirstProof (Abouzaid et al., 2026a), to avoid imposing formatting requirements that could divert too much effort from attempting the problem.

If the conjecture was marked unresolved after the first pass, GPT-5.6-Sol Pro was prompted to resolve the conjecture. To do this, it was given the conjecture with necessary context (consisting of definitions, notation, and relevant statements required to understand the conjecture), along with the exact output of the first pass, and instructions to produce a proof or counterexample. If the conjecture was still unresolved after the second pass, we prompted GPT-6-Astra Pro to resolve the conjecture using the same input as the second pass, along with the second pass response. This pass was motivated by our initial review of second pass attempts, in which it seemed the models made nontrivial progress. The third pass was included to test whether this progress in the second pass would yield a complete proof with a stronger model. Web search was permitted in the first pass, but we removed it for the second and third pass after initial review of first pass responses showed examples that seemed more focused on literature review than thinking about the problem.

5 Results and Analysis

In total, the three passes produced 70 counterexamples and 55 proofs to the 200 conjectures. Not all of these stood up to LLM-based verification. Figure 2 shows the overall results, both by math subject classification and by the pass in which a resolution was proposed. Among the counterexamples, 49 were found in the first pass, 15 in the second, and 6 in the third. Among the proofs, 33 were found in the first pass, 11 in the second, and 11 in the third. The number of proposed counterexamples outnumber proofs approximately 60/40 in the first two passes, but by the third pass, the number of

proposed proofs outnumber counterexamples. In this section, we give an overview of results, discussing the LLM correctness review, our findings about the quantity and nature of citations provided by the solver models, and the length and cost of responses, before discussing the counterexamples and proofs in more detail.

LLM review for correctness For the proposed resolutions, we asked a new instance of the same model to verify the correctness of the attempt. The reviewer model was given the proposed proof or counterexample and the full paper. The prompt used can be found in Appendix C.1. All third pass proofs and counterexamples, and between 80 and 85% of first and second pass proofs and counterexamples, passed this check. We reviewed the verification to check for obvious issues (such as a verifier run refuting a proof with an edge-case counterexample that was likely excluded by the authors), but true verification would require expert review that the proposed counterexample or proof was settling the statement intended by the authors, and is outside of scope for our study. Our analysis of proofs and counterexamples in the sections below focuses on just the ones that passed this validation check.

Resolution citations and other references In FirstProof Second Batch (Abouzaid et al., 2026b), the authors noticed models often failed to appropriately attribute results. This motivated our decision to explicitly request citations in the first pass prompts. This was effective at getting models to include citations. Under a broad definition of citations as an explicit reference to an external work, theorem, named result, etc, we found that the first pass responses included an average of 3.3 citations and a median of 2 distinct citations. (This count does not including the arXiv preprint which contained the conjecture, which was cited in 93% of responses.) Responses that were unresolved had the most citations; counterexamples were most likely to have no citations other than the source paper. These citations were most often arXiv preprints (29%) or journal articles or other papers (24%), but also included named results (20%) and constructions, formulas, or similar (11%) (see Figure 9 in Appendix B).

The second and third pass responses also included citations under this broad definition (both had a median of 2 distinct citations, with means of 2.8 and 2.2, respectively) but the type of citations were very different compared to the first pass responses: about 60% were named theorem or results, and the remainder were different sub-categories of constructions, formulas, or other criteria. The prompt the second- and third- pass solvers received included instructions not to use web search or discuss the literature status of the conjecture, and it's possible that (i) the distinction between giving proper attribution and discussing the literature status is subtle and (ii) not being able to use web search and check the accuracy of a citation makes a model less likely to include a citation.

Requesting citations in the first pass came with a trade-off: responses seemed to be more focused on literature review/search and discussion of the status of the conjecture rather than working on the problem. We hypothesized that specifically requesting precise citations and providing web search exacerbated this issue. Indeed, the models used web search more than we expected: in the first pass, the median number of web search calls per conjecture was 18 and the median number of web search queries was 40. (The distribution of web search calls is shown in Figure 8 in Appendix B.) Accordingly, the median number of input tokens in this pass was 131,783, as opposed to 6,309 in the second pass.

Response length We measured the length of responses in two ways: by the prose word count, and the total number of claims with proof (Figure 3). The third-pass responses are the longest by both measures. Unresolved responses tend to be longer in word count compared to proofs and counterexamples but less so in the second pass. The prose word count is shown by subfield in Appendix B; there seems to be more variability in response length within the subfields than between subfields. In contrast to the number of lemmas, theorems, propositions, and corollaries, responses contained very few definitions that were introduced by the solver (as opposed to definitions from the source paper, external literature, or standard mathematical background).

Cost Plots showing input and output token usage for the first, second, and third passes are shown in Appendix B. The median number of input tokens/output tokens was 131,783/24,956, 6,309/34,522, and 9,727/29,022 for the first, second, and third passes, respectively, so the median cost by pass was approximately $0.97, $0.41, $1.71 at the time of writing.

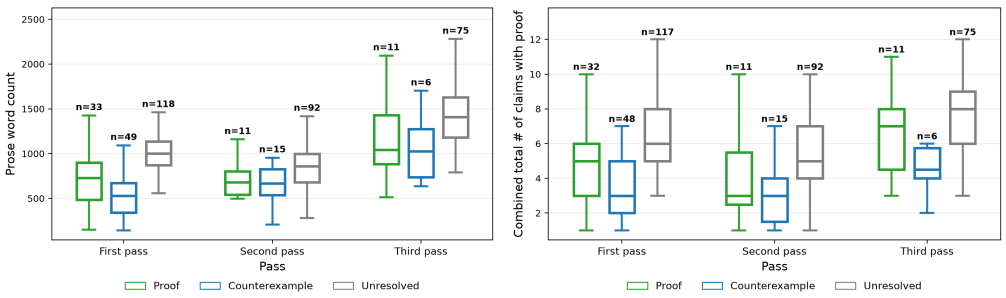

Figure 3: Response length is measured by the prose word count (left) and the total number of claims with proof (right).

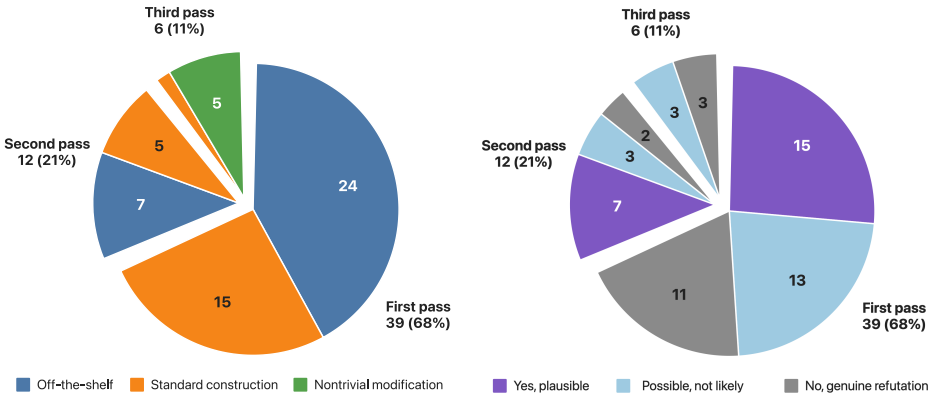

Figure 4: Left: LLM classification of object novelty for counterexamples that were ‘verified‘ by the review pass. Counterexamples are classified as ‘off-the-shelf’, ‘standard construction’, or ‘nontrivial modification’. Right: LLM responses to the question ‘is it plausible that the conjecture’s authors implicitly excluded this case’ (i.e., the conjecture is misstated rather than genuinely refuted)?

5.1 Counterexamples

Our review of counterexamples is focused around two questions:

(i) How often was a conjecture refuted by an “off-the-shelf” counterexample, meaning an example a specialist in the field would have on their standard list of things to try?

(ii) How often is the counterexample provided implicitly excluded by the authors?

The motivation for the first question comes from the fact that in many fields of mathematics, there are “standard” counterexamples documented in books like Counterexamples in Topology (Steen et al., 1978) and Counterexamples in Analysis (Gelbaum & Olmsted, 2003). We were also influenced by a recent blog post by Timothy Gowers discussing the kinds of mathematics problems LLMs are good at, which listed different methods for counterexample search (Gowers, 2026). Furthermore, in our initial manual review of counterexamples, we noticed many straightforward counterexamples in the first pass responses.

This observation also motivated question (ii) since this simplicity was extreme in certain cases and it seems likely that it was unaccounted for by the authors (even if their conjecture framing was not precise enough to exclude these). For example, counterexamples to some of the conjectures in combinatorics included the cycle graph on five vertices and the complete graph on five vertices. Another example, given to a number theory conjecture, was the trivial group. The model responses themselves also seemed to suggest that the provided counterexample may have been implicitly excluded by the authors. Specifically, we noticed counterexamples sometimes included a paragraph titled “Scope”, with language similar to the following: “The counterexample refutes the conjecture

exactly as stated." and then go on to list an amendment of the conjecture that the provided counterexample did not address. We hypothesize that at least in some cases this amendment was what was intended by the paper authors. In another example, the model labeled its counterexample "literal counterexample": a single-vertex graph.

5.1.1 METHODS

After a manual review of the counterexamples, we used an LLM review to produce the results shown in Figure 4. For each counterexample, we prompted a new instance of the same model that produced the counterexample to classify the object as *off-the-shelf* (an example a specialist would have on their standard list of things to try), a *standard construction* (obtainable from off-the-shelf examples by standard modifications within the subfield), a *nontrivial modification* (built from known ingredients but requiring a non-obvious choice), or *genuinely novel* (introducing an object or construction not reducible to known ingredients). We also asked the model to give an assessment of the *novelty of the argument*. This asked about the difficult of the argument that the provided object is indeed a counterexample. The review prompt also asked the reviewing model to assess whether it is plausible that the authors of the conjecture implicitly excluded this counterexample and/or the conjecture was too imprecise.

The complete prompt, which included several other questions including the overall novelty, is given in Appendix D.

5.1.2 RESULTS

Novelty of counterexamples Figure 4 (left), which show LLM judgements of novelty for the verified counterexamples, agree with informal observations we made during manual review. Both the first pass and second pass counterexamples are about 60% "off-the-shelf", and the remainder were all classified as "standard construction." In contrast, most third pass counterexamples were given ratings of "nontrivial modification" (5/6). The third pass counterexample responses were also longer than counterexample responses from the first two passes, in both word count and the combined total number of claims proved (see Figure 3), and these verifications were more often called out as nonroutine by the LLM reviews.

Counterexamples exploiting missing assumptions or mistakes Figure 4 (right) shows the fraction of counterexamples that a reviewer judges to be the result of imprecision in the conjecture formulation. This makes up a significant portion of first-pass and second-pass counterexamples. Some excerpts from model reviews which are labeled as "very plausibly" due to imprecision include:

- "This appears to expose a convention mismatch rather than defeat the intended representation-theoretic conjecture."
- "The counterexample therefore diagnoses a misstatement or order-of-operations error rather than overturning the intended linearized double-shuffle conjecture."
- "Consequently, this looks more like exploitation of a missing projectivity/properness hypothesis than a counterexample to the intended geometric conjecture"

In summary, the results suggest that while some counterexamples are off-the-shelf or the result of imprecision, some likely represent genuine novelty. However, we wish to emphasize that our overall results are also reflections of the data in addition to GPT-5.6-Sol and GPT-6-Astra capabilities. That is, one can't simply say 'Out of 200 conjectures, there were 70 proposed counterexamples, 58 which held up to automated verification' without the taking into consideration that

- Many mathematical conjectures are not self-contained statements, and rely on context in other portions of the paper.
- GPT-5.6-Sol may interpret extracted conjecture statements too literally; mathematicians working in the same subfield may better be able to apply common or implicit assumptions.
- A random sample of conjectures from arXiv papers (preprints that have not yet undergone peer review) is likely to contain conjectures that require small corrections.

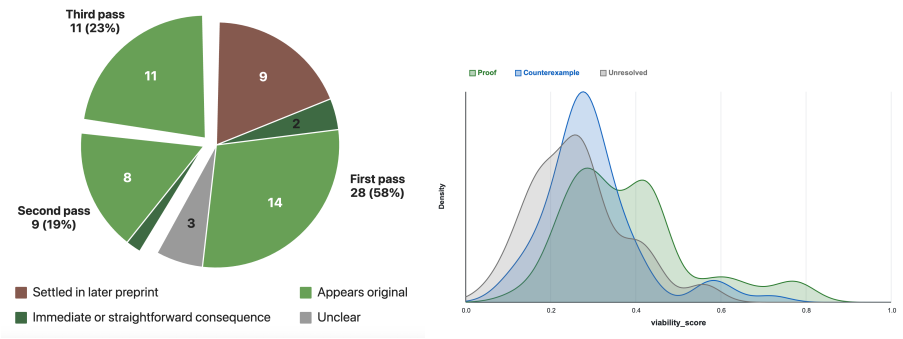

Figure 5: Left: Relation of proof responses to the existing literature. Right: GPT-5.4 scores of viability were weakly predictive of outcome.

One excerpt from an overall verdict portion of an LLM review felt particularly representative of many of the first and second-pass attempts. "It is a useful correction to the conjecture, but it contains neither a novel object nor a novel argument."

5.2 Proof analysis and progress across attempts

In this section we focus on two aspects of the proofs and unresolved responses: their relationship to existing results, and the extent to which later attempts used the results and strategies developed in earlier passes.

Relationship to existing literature In our initial review of proofs from the first pass, we found that a surprising number of conjectures had been resolved in a revision of the source preprint or a subsequent preprint by different authors[2]. Still others appeared to be fairly immediate consequences of existing results. Based on this review, we classified proofs into the following categories:

(1) Conjectures that had been settled in subsequent preprint(s).

(2) Proofs that were immediate or straightforward consequence of other results.

(3) Proofs that appear original in our non-specialist review

(4) `Unclear`, which was given when the arguments used external results but we were not able to determine whether the additional reasoning was routine enough to fall into category (2).

Figure 5 shows the classification by pass. We assigned labels manually, using the responses and relevant portions of the LLM verification review. Interestingly, we found that 3 out of 9 of the proofs that were settled in a later preprint included an independent argument. So even when a model response began by identifying that the conjecture had been proven, it didn't always simply reproduce the proof.

Apparent progress in unresolved attempts Most striking about the unresolved attempts was how much progress they appeared to make. Figure 3 shows the total number of statements with proof (theorems, lemmas, propositions, and corollaries) by pass and by classification of attempt (proof, counterexample, or unresolved). The median number of statements with proof in the first, second, and third pass unresolved attempts were 6, 5, and 8 respectively. Reading the responses gives the impression that one has almost reached a resolution, but of course it is not actually clear whether the attempt is useful for solving the conjecture or even correct.

We were interested in identifying response characteristics predictive of solver resolution in a future pass. We found that both prose word count and the count of number of claims with proof did not predict next-pass success: responses that remained unresolved in the next pass typically had comparable or higher median lengths than responses that were resolved in the next pass. We also checked whether conjecture viability scores were predictive of overall outcome (recall that the conjectures

[2]These cases illustrate the limitations of the literature-status filtering used in the dataset construction.

in the OpenConjecture dataset are scored by interestingness and viability). We found a statistically significant difference: the mean viability score of conjectures for which the model produced a proof was 0.34 compared to 0.26 for conjectures that remained unresolved. However, there is substantial overlap between the groups (Figure 5).

Later attempts sometimes identified issues with earlier attempts. For example, one third-pass response stated "One proposed route in both preceding attempts can be ruled out". Another reported that the "core perimeter inequality" in the second attempt was false. These are assessments from the model and not independent expert assessments, but still illustrate the challenge of judging progress towards a proof or counterexample.

Did third-pass proofs build on earlier work? To obtain a more concrete indication of the usefulness of earlier work, we checked whether the third pass proofs actually used the apparent partial progress from the first two passes, or took a different approach. We found in 7 of the 11 proofs, the third pass proof builds on the first two passes (without attribution/acknowledgement that the results being reproduced were provided in previous passes). In the other four examples, the model changed approach from the approach taken in the first two passes.

6 LIMITATIONS

Correctness of proposed resolutions Our results discuss candidate resolutions and outcomes of LLM review, not expert-verified solutions. Each proposed resolution was reviewed by a new instance of the same model that produced it. Consequently, the solver and reviewer model may share mathematical errors. Our manual review is not a substitute for specialist review or formal verification. As discussed, we suspect some of the proposed solutions may not refute the conjecture intended by the authors. **The reported numbers of proposed proofs and counterexamples and review-pass rates should not be interpreted as rates of newly solved open problems.**

Scope of the evaluation We evaluated three GPT models, rather than conducting a controlled comparison across model families or seeking the best achievable performance of each model by running multiple attempts or evaluating the models with their recommended harnesses. Future work would be needed to assess run-to-run variability or how alternative prompts and tools would change the results.

7 DISCUSSION

We studied the outcomes of applying frontier GPT models to conjectures drawn from recent arXiv math preprints, rather than problems carefully selected for a benchmark. Our findings have implications for those interested in creating datasets from arXiv literature: just because a problem is from a recent paper and passes automated open-status filtering does not mean it remains unresolved or sufficiently difficult for a benchmark. Our work also highlights the difficulty of extracting the relevant context required for a standalone statement of a mathematical conjecture and the need for specialist review of proposed counterexamples.

Overall, our analysis emphasizes the different dimensions of evaluating a mathematical proof or counterexample: correctness, fidelity to the intended statement, relationship to existing literature, and the mathematical contribution (which itself could likely be broken into further dimensions). Evaluation that discusses these dimensions would help increase understanding of model capabilities and their value to mathematical research.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the research methodology and to help write prompts that were used in our study. Additionally, we used generative AI tools to create figures and tables for our paper, write code, and give feedback on the readability and clarity of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We have made the following efforts to ensure reproducibility: (1) the OpenConjecture dataset is publicly available, (2) the solver model responses will be made publicly available, to allow for reproducibility of the analysis, and (3) the prompts we used are included in the Appendix.

References

Mohammed Abouzaid, Andrew J Blumberg, Martin Hairer, Joe Kileel, Tamara G Kolda, Paul D Nelson, Daniel Spielman, Nikhil Srivastava, Rachel Ward, Shmuel Weinberger, et al. First proof. *arXiv preprint arXiv:2602.05192*, 2026a.

Mohammed Abouzaid, Nikhil Srivastava, Rachel Ward, and Lauren Williams. First proof second batch, 2026b. URL https://arxiv.org/abs/2606.18119.

Andrei Balakin, Miklós Bóna, Marie-Charlotte Brandenburg, Clara Briand, Veronica Calvo Cortes, Shelby Cox, Jesus A De Loera, Danai Deligeorgaki, Hannah Friedman, Tim Gehrunger, et al. Benchmarks in leipzig. *arXiv preprint arXiv:2606.05818*, 2026.

Jasper Dekoninck, Tim Gehrunger, and Martin Vechev. ArXivMath: Evaluating LLMs on mathematical research problems from recent ArXiv papers. https://matharena.ai/arxivmath/, Feb 2026. MathArena.

Suhaas Garre, Erik Knutsen, Sushant Mehta, and Edwin Chen. Riemann-bench: A benchmark for moonshot mathematics. *arXiv preprint arXiv:2604.06802*, 2026.

Bernard R Gelbaum and John MH Olmsted. *Counterexamples in analysis*. Courier Corporation, 2003.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, and Emily de Oliveira Santos. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Timothy Gowers. What sort of maths are LLMs good at? Gowers's Weblog, August 2026. URL https://gowers.wordpress.com/2026/08/12/what-sort-of-maths-are-llms-good-at/. Accessed: 2026-09-24.

Haocheng Ju and Bin Dong. AI for mathematics: Progress, challenges, and prospects. *arXiv preprint arXiv:2601.13209*, 2026.

Seewoo Lee. Awesome AI for math: A curated list of works that use AI for mathematical discoveries. https://seewoo5.github.io/awesome-ai-for-math/, 2026. Accessed: 2026-09-01.

OpenAI. An OpenAI model has disproved a central conjecture in discrete geometry. OpenAI Blog, May 2026. URL https://openai.com/index/model-disproves-discrete-geometry-conjecture/.

OpenAI. Finite time blowup for Navier–Stokes. Preprint, September 2026. URL https://cdn.openai.com/pdf/32d9f210-8b73-45e0-91bc-82a30aef8a9a/navier-stokes.pdf.

OpenAI. Ten advances in mathematics and theoretical computer science. OpenAI Blog, August 2026. URL https://openai.com/index/ten-advances-in-mathematics/.

Antoine Peyronnet, Fabian Gloeckle, and Amaury Hayat. LemmaBench: A live, research-level benchmark to evaluate LLM capabilities in mathematics. *arXiv preprint arXiv:2602.24173*, 2026.

Johannes Schmitt, Gergely Bérczi, Jasper Dekoninck, Jeremy Feusi, Tim Gehrunger, Raphael Appenzeller, Pieter Belmans, Alessio Bottini, Jim Bryan, João Camarneiro, et al. IMProofbench: Benchmarking AI on research-level mathematical proof generation. *arXiv preprint arXiv:2509.26076*, 2025.

Avi Sircar. Mathematicians grapple with a 'very rapid and very unsettling change' as AI cracks yet another century-old problem. *Fortune*, July 2026. URL https://fortune.com.

Guijin Son, Seungone Kim, Catherine Arnett, Hyunwoo Ko, Hyein Lee, Hyeonah Kang, Jiang Longxi, Jin Yun, JungYup Lee, Kyungmin Lee, et al. SOOHAK: A mathematician-curated benchmark for evaluating research-level math capabilities of LLMs. *arXiv preprint arXiv:2605.09063*, 2026.

Lynn Arthur Steen, J Arthur Seebach, and Lynn A Steen. *Counterexamples in topology*, volume 18. Springer, 1978.

A OpenConjecture dataset descriptive statistics

Labeling The specific criteria given to GPT-5 Mini for labeling are as follows:

- `real_open_conjecture`: the exact extracted statement remains unresolved as written and is being posed as an active open target."
- `not_real_conjecture`: the exact extracted statement is already proved/refuted/-known, is just a solved special case of a broader parent conjecture, is only background motivation or survey context, is inactive/commented/empty, or follows immediately from known results.
- `uncertain`: insufficient context.

In the labeling pass, GPT-5 Mini is given the conjecture block and 20 lines before and after the conjecture block.

The June 21st snapshot consists of 3030 conjectures labeled `real` by GPT-5 Mini. (Total, there were 3710 conjectures. 689 were labeled as not real, and 19 as uncertain). The `real` conjectures come from 1675 papers. The number of conjectures per paper (of the papers with at least one conjecture) is heavily right-skewed, with the vast majority of conjectures coming from papers with one or two conjectures. (See Table 1).

769 of these conjectures came from 432 papers with a non 'math.*' primary category: 95 were cross-listings, and 337 had an empty primary category. Most of these 337 papers appeared to be updates, based on their published at range.

Table 1: Per-paper conjecture count distribution over 1,675 papers

Conjectures per paper	1	2	3	4	5	6	7	8	9	10	11	14	15	18	20	26	27	36
Paper count	1049	353	130	66	30	12	9	5	6	6	1	1	2	1	1	1	1	1

B Additional results

Figure 7 shows the distribution of input tokens (left) and output tokens (right) by pass. Note that the first pass was the only pass to allow web search, which counts towards input tokens. The distribution of the number of web search calls per response is shown in Figure 8.

Figures 9 through 11 show the breakdown of citations provided in model responses.

C Conjecture solution pipeline: Additional details

Providing models with the full paper versus extracting selected context For the first pass, we considered whether to give the model the full paper or to give it just the conjecture and relevant context. In the second and third pass, we wanted the model to focus on building on the previous attempt(s), so we gave the model the conjecture and relevant context rather than the full paper, along with the previous attempts.

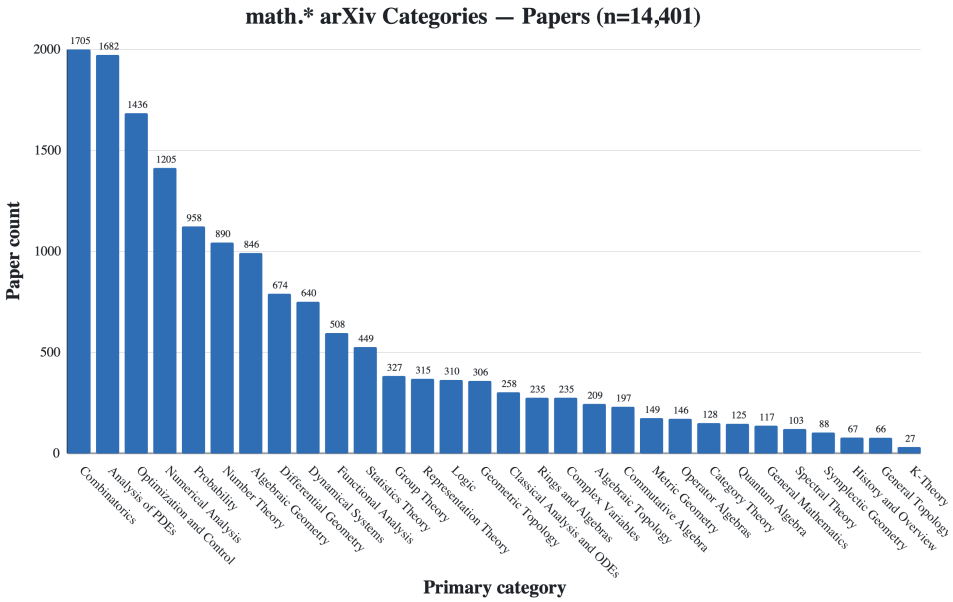

Figure 6: math.* category of papers from OpenConjecture June 21st snapshot.

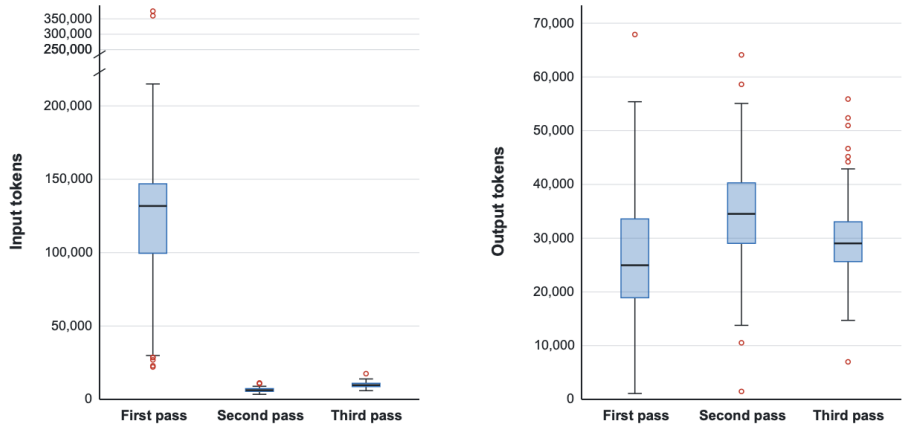

Figure 7: Left: distribution of input tokens per solver pass. First pass (GPT-5.6-sol): N=197 (3 of 200 CIDs had missing token data), second pass (gpt-5.6-sol): N=118, third pass (GPT-6-astra): N=92. Right: distribution of output tokens per solver pass.

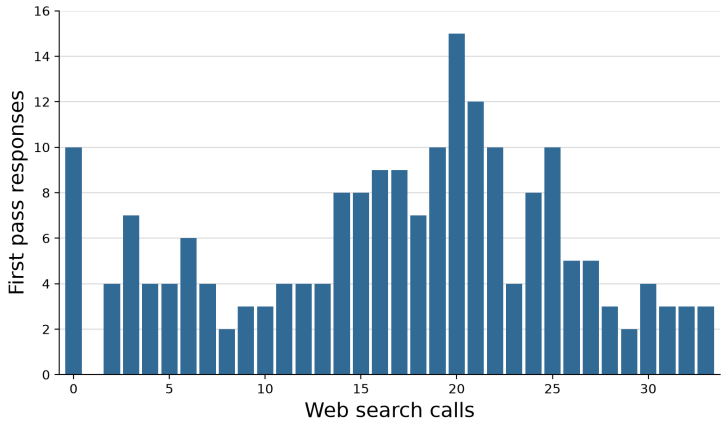

Figure 8: Distribution of the number of web search calls per response in the first pass.

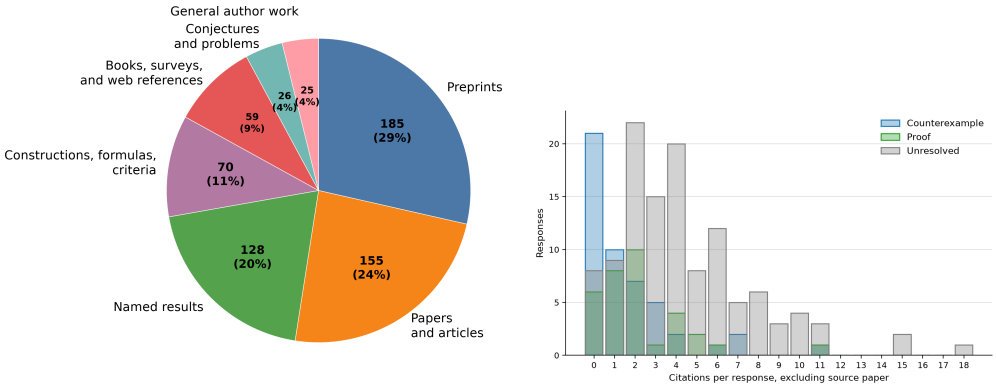

Figure 9: Left: Taxonomy of citations in first pass responses. Right: distribution of number of citations per response, excluding citations of the source paper.

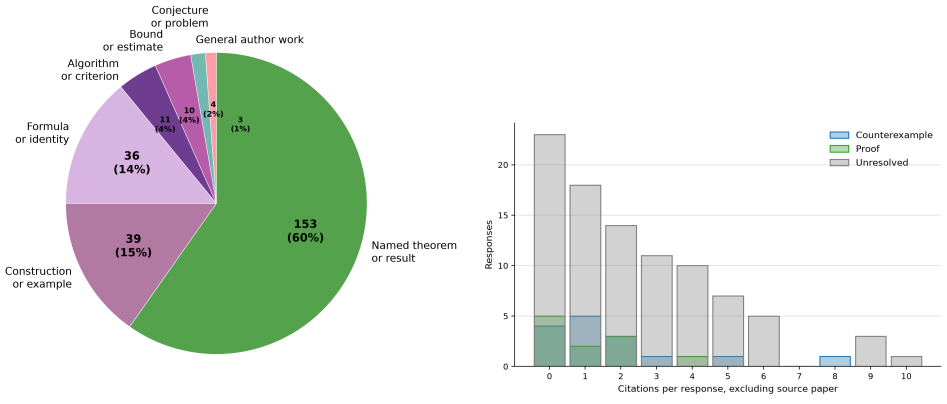

Figure 10: Left: Taxonomy of citations in second pass responses. Right: distribution of number of citations per response, excluding citations of the source paper.

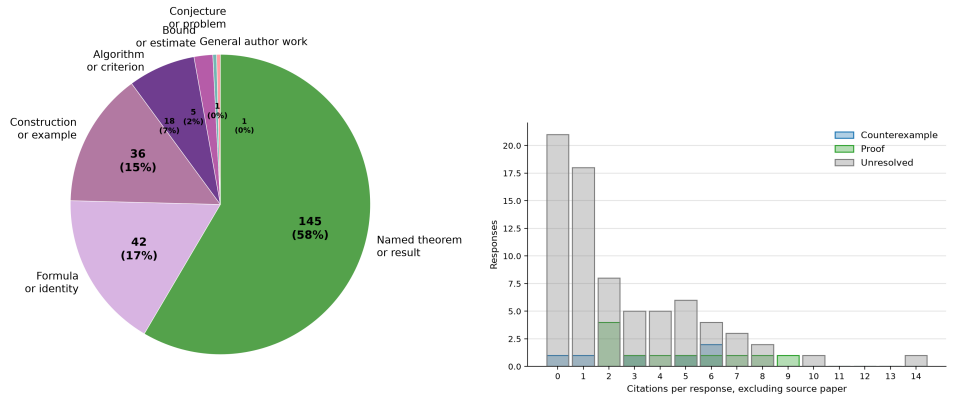

Figure 11: Left: Taxonomy of citations in thid pass responses. Right: distribution of number of citations per response, excluding citations of the source paper.

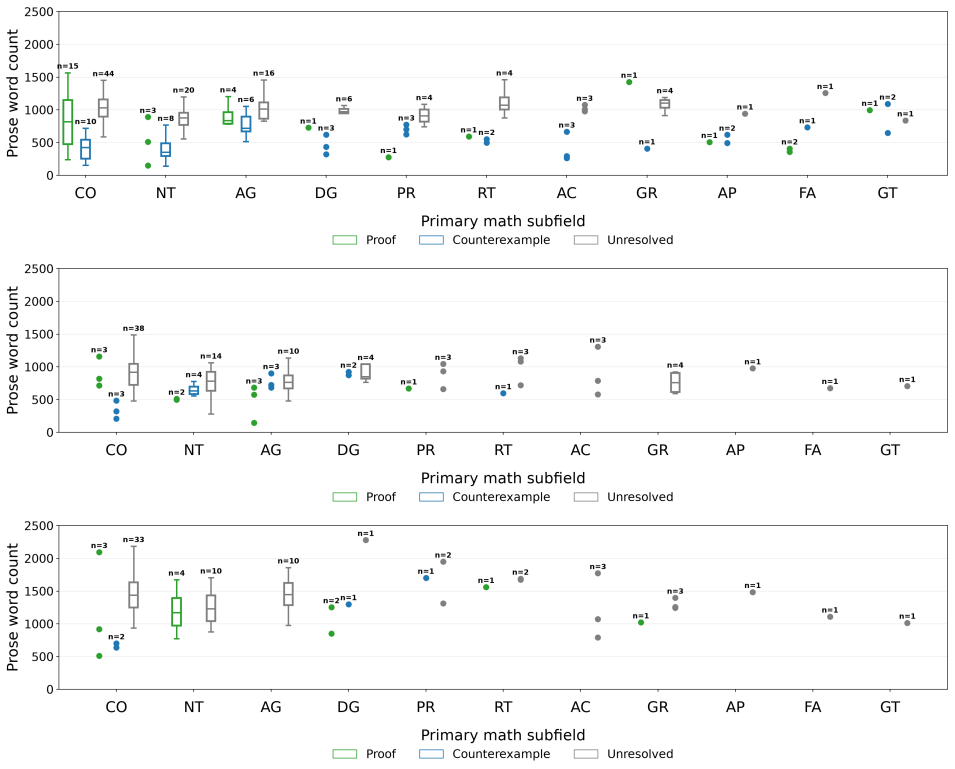

Figure 12: Prose word count by math.* classification.

# defns	First pass			Second pass			Third pass			Total
	Proof	CEX	Unres.	Proof	CEX	Unres.	Proof	CEX	Unres.	
0	28	38	89	9	11	82	10	4	69	340
1	3	5	25	1	0	6	0	1	2	43
2	1	3	3	0	2	1	1	0	3	14
3	0	1	1	0	0	2	0	1	0	5
4	1	2	0	1	2	0	0	0	0	6
5	0	0	0	0	0	1	0	0	0	1
6	0	0	0	0	0	0	0	0	1	1
Total	33	49	118	11	15	92	11	6	75	410

Table 2: Distribution of explicit solver-introduced definitions by solver pass and response type. Counts exclude definition items classified as from the source paper, external literature, or standard mathematical background; retained items are definitions, notation, conventions, interpretations, or auxiliary constructions introduced by the solver response itself.

Extracting relevant context To create a self-contained conjecture statement, we used GPT-5.6-Luna to extract the relevant context from the paper (using the prompt in Appendix C.2 below). However, even after iterating on this prompt, and experimenting with stronger models, these packs were often deemed to contain insufficient context by an LLM review. It is difficult to reliably extract all of the implied assumptions, which are often stated in the introduction or can even be omitted (but may be clear to others working in the subfield).

C.1 SOLVER PROMPTS

System prompt used in the first pass

```
Carefully consider the following research-level mathematical
    conjecture and make a sustained attempt to prove or refute it.
Do not assume it is true because it is stated as a conjecture. Use the
    context provided to interpret the conjecture in the spirit that
    the authors intended.
You may use web search.

Your response must reach exactly one verdict, stated on the first line
    as `VERDICT: PROOF`, `VERDICT: COUNTEREXAMPLE`, or `VERDICT:
    UNRESOLVED`:
1. PROOF: a complete, rigorous proof, organized into lemmas and
    theorems, with every nontrivial step justified.
2. COUNTEREXAMPLE: an explicit object together with a complete
    verification that it violates the conjecture.
3. UNRESOLVED: the strongest results you have actually established,
    and a precise statement of the remaining gap.

Standards:
- UNRESOLVED is an acceptable outcome. A claimed proof or
    counterexample with a gap is far worse than an honest UNRESOLVED.
    Before finalizing, re-check every step; if any step depends on a
    claim you have neither proved nor cited, downgrade to UNRESOLVED.
- Do not invent definitions, theorem statements, citations, or
    computational results. Cite every external result you rely on
    precisely (authors, title, and the exact statement used).
- If the conjecture or a close variant is already settled in the
    literature, say so, cite the source, state whether it is
    essentially identical or merely related, and explain how it
    settles (or fails to settle) the conjecture as stated.
- If the statement is ambiguous or missing context, resolve the
    omission using available tools, or state your interpretation
    explicitly.
```

```
Write the mathematical content in LaTeX using theorem/lemma/proof
    environments. A separate pass will produce the final compilable
    document.

The target conjecture and the source material supplied follow in the
    user message.
```

System prompt used in the second pass.

```
You are an experienced mathematician in a research pipeline.

You are given:
1. a research-level mathematical conjecture and mathematical context
    relevant to the conjecture; and
2. a previous solution attempt containing the interpretation and
    literature status, candidate lemmas and propositions, and a
    statement of the remaining gap.

Your primary objective is to construct a complete rigorous proof or an
    explicit counterexample to the target statement (taking into
    consideration the context, in the spirit that the authors
    intended), using the previous attempt as a starting point.

This is a closed-book mathematical research run. Do not use web
    search, discuss the literature status of the conjecture, or infer
    that it is unsolvable because it is labeled open. Treat the
    previous attempt as untrusted research notes: verify its relevant
    claims before relying on them, and do not inherit its verdict.

Make a sustained attempt. Do not stop because the problem is labeled
    open or because the first approach fails.
Isolate the smallest promising remaining lemma and attack it directly.
If it reduces to another lemma, attack that lemma as well; if one
    route fails, try a materially different route.

Structure the mathematical argument using careful theorem, lemma, and
    proof environments, with LaTeX notation. A separate pass will
    produce the final compilable document.
```

The prompt for the third pass was almost identical to the second pass, except for the description of what is being provided. The model is instructed that it is given "(2) a first-pass attempt: an initial exploratory attempt (which had web search available) containing the interpretation and literature status, candidate lemmas and propositions, and a statement of the remaining gap; and (3) a second-pass attempt: a closed-book follow-up attempt that took the first-pass attempt as its starting point and made further progress."

C.2 System prompt for context extraction

System prompt used to extract relevant conjecture context.

```
You assemble interpretation context for a mathematical conjecture.

Your task is only to identify the definitions, notation, applicable
    standing assumptions, and directly referenced results needed to
    interpret the exact statement. Do not solve, prove, refute, or
    evaluate the conjecture.

Rules:
```

```
1. Select span IDs exactly as displayed. Do not add brackets, change
    line ranges, merge ranges, or invent IDs.
2. Include definitions and notation for every nonstandard symbol and
    technical term, including transitive definitions needed to
    understand them.
3. Include only assumptions whose scope applies at the conjecture.
4. A sentence saying "throughout this section" does not apply in
    another section.
5. Results that are merely topically related are not interpretation
    context.
6. If an indispensable definition exists only in an external citation,
    report it as an external dependency.
7. If a dependency cannot be resolved, report it as unresolved.
8. The dependency "meaning" fields are concise navigation summaries.
    The source spans remain authoritative.
9. Do not select the conjecture statement itself.
10. For dependencies[].source_span_id and
    standing_assumptions[].source_span_id,
    use null or an ID that you also return under selected_spans. These
    fields identify support inside the included context; they do not
    add content to the pack. If no selected span supports it, use null
    (and report a genuine gap under unresolved_dependencies or
    external_dependencies).
11. Order selected_spans by mathematical dependency: prerequisites
    should have smaller dependency_order values than objects defined
    from them.
12. Classify each dependency's resolution:
    - included_source: defined by an included source span;
    - statement_bound: introduced or constrained directly by the
    conjecture;
    - standard: standard mathematical notation or terminology, OR a
    well-known published mathematical object, theorem, or term that a
    competent mathematician recognizes without a definition (e.g. ring
    class field, quadratic order of a discriminant, class number,
    ideal class group, Galois group, standard L-functions). Such
    objects need not be defined in the paper to be `standard`;
    - external_required: a definition or result specific to a cited
    external source, indispensable to interpret the target, AND not
    part of standard mathematical knowledge. Do NOT use this for
    well-known published objects - those are `standard`;
    - unresolved_internal: missing but expected elsewhere in this
    paper;
    - unresolved_unknown: unresolved with unknown location.

Return exactly one JSON object between BEGIN_JSON and END_JSON:

{
  "statement_self_contained": "yes|no|partial",
  "paper_context_sufficient": "yes|no|partial",
  "selected_spans": [
    {
      "span_id": "exact candidate id",
      "role": "definition|assumption|result|orientation",
      "dependency_order": 1,
      "reason": "why this span is needed"
    }
  ],
  "dependencies": [
    {
  "term": "symbol or technical term",
  "kind":
    "symbol|technical_term|quantifier_domain|reference|assumption",
  "meaning": "concise source-grounded meaning",
```

```
  "resolution":
    "included_source|statement_bound|standard|external_required|unresolved_internal|unresolve
  "source_span_id": "exact candidate id or null"
}
  ],
  "standing_assumptions": [
    {
      "text": "concise description",
      "source_span_id": "exact candidate id",
      "scope": "global|section|subsection|environment|unknown",
      "applies": true
    }
  ],
  "conflicts": [],
  "external_dependencies": [],
  "unresolved_dependencies": []
}

Use double-quoted JSON strings and no Markdown fences.
```

D LLM REVIEWER PROMPTS

D.1 COUNTEREXAMPLES

Prompt used for counterexample review.

```
You are an experienced mathematician auditing a proposed
    counterexample. You are given:

1.  a research-level mathematical conjecture and mathematical context
    relevant to the conjecture; and
2.  a proposed counterexample.

Your task is to independently determine whether the proposed witness
    genuinely refutes the conjecture. Treat the supplied derivation as
    untrusted. Check the hypotheses, instantiate the conjecture and
    its exact negation at the witness, and recompute the decisive
    quantities from the original definitions. Begin with exactly one
    verdict:
- `VERIFIED_COUNTEREXAMPLE`: all hypotheses hold and the negation of
    the conjectured conclusion is rigorously established.
- `INVALID_COUNTEREXAMPLE`: a required hypothesis fails, or the
    conjectured conclusion is rigorously shown to hold for the witness.
- `INCONCLUSIVE`: neither verdict is rigorously established. A flaw in
    the proposed derivation alone does not make the counterexample
    invalid.

After the verdict, give concise mathematical reasoning identifying the
    witness, the relevant hypothesis checks, the required negation,
    and the decisive certificate. State any computational or heuristic
    dependencies. Use LaTeX notation. A separate pass will produce the
    final compilable document.
```

Prompt used for counterexample novelty assessment.

```
You are an experienced mathematician reviewing assessing the NOVELTY
    of a proposed counterexample to a conjecture. You are given:
        1. a research-level mathematical conjecture and mathematical
    context relevant to the conjecture; and
```

```
    2. a proposed counterexample.
Assume the counterexample is correct. Do NOT verify it;
correctness is assessed separately. Your task is only to judge how
novel it is.
You have access to web search. Use it to ground your assessment:
check whether the object or construction has a name, appears in
standard references, or has been discussed in connection with this
conjecture. Search before classifying, not only when asked for
references.

Answer the following.

A. KEY IDEA. In one or two sentences, identify the essential
ingredient of this counterexample: what object is used, and what
feature of it makes the conjecture fail?

B. NOVELTY OF THE OBJECT. Classify the object/construction into
exactly one of:
(1) OFF-THE-SHELF: an example a specialist would have on their
standard list of things to try (e.g. the first few things one
checks when testing any statement in this area).
(2) STANDARD_CONSTRUCTION: not itself off-the-shelf, but
obtainable from off-the-shelf examples by standard construction
methods in the subfield without any new idea.
(3) NONTRIVIAL MODIFICATION: built from known ingredients but
requires a non-obvious choice, tuning, or combination that a
specialist would not find by routine search.
(4) GENUINELY NOVEL: introduces an object or construction not
reducible to known ingredients.
Explain your classification, and name the off-the-shelf examples
or standard constructions it resembles or is built from.

C. NOVELTY OF THE ARGUMENT. Is the argument that this object
violates the conjecture itself routine once the object is written
down, or does the difficulty lie in the verification?

D. PRIOR APPEARANCE. Address separately:
(i) Has this object or construction appeared in the literature
(for any purpose)?
(ii) Has it previously been observed to be relevant to this
conjecture or closely related statements?
(iii) Is it plausible that the conjecture's authors implicitly
excluded this case (i.e., the conjecture is misstated rather than
genuinely refuted)?
For any claimed reference, give the type of source (paper,
textbook, survey, folklore, online discussion) and your confidence
(HIGH / MEDIUM / LOW). Do NOT invent citations. You can use web
search as needed.

E. OVERALL VERDICT. One of: NOT NOVEL / MARGINALLY NOVEL / NOVEL /
UNRESOLVED, with a one-paragraph justification. Use UNRESOLVED
only when you genuinely cannot decide, and state what information
would resolve it.

End your response with a line of the exact form `NOVELTY_VERDICT:
NOT_NOVEL`, `NOVELTY_VERDICT: MARGINALLY_NOVEL`, `NOVELTY_VERDICT:
NOVEL`, or `NOVELTY_VERDICT: UNRESOLVED`.
```

D.2 PROOFS

Prompt used for proof review.

```
You are an experienced mathematician auditing a proposed proof.

        You are given:
        1. a research-level mathematical conjecture and the full paper
    containing the conjecture; and
        2. a proposed proof of the conjecture.

        Your task is to independently determine whether the proposed
    proof rigorously establishes the conjecture. Treat the supplied
    derivation as untrusted. Verify the argument from the original
    definitions and accepted results. Independently check every
    implication and claimed equivalence,
        and every use of an external theorem, including whether all of
    its hypotheses hold and whether its conclusion is strong enough.
    Look for hidden assumptions, circular reasoning, missing cases,
    and gaps in implications.

        You may fill in minor routine details, but you must explicitly
    say when you have done so. Do not silently repair substantive gaps.

        You have access to web search. Use it when the proposed proof
    depends on a specialized external result whose exact statement,
    hypotheses, definitions, or version are not fully supplied.

        If the proposed proof says that the conjecture was settled in
    a revision or later work, state whether it relies on that result
    or provides an independent proof. If it relies on the result,
    verify the cited statement, its hypotheses, and that it directly
    implies the conjecture as stated. If it provides an independent
    proof, verify that argument independently; the existence of
    another proof is not evidence that the proposed proof is valid.

        Begin with exactly one verdict:

        - `VERIFIED_PROOF`: The submitted argument proves the
    conjecture, with at most explicitly identified minor routine
    completions.
        - `INVALID_PROOF`: The submitted argument does not establish
    the conjecture because it contains a false or unjustified
    substantive step, circular reasoning, a missing necessary case, an
    unmet hypothesis, or a proof of a weaker or different statement.
    This verdict does not imply that the conjecture is false.
        - `DISPROVED_STATEMENT`: You have rigorously established a
    counterexample to the conjecture as stated. Verify that the
    counterexample satisfies every hypothesis and violates the claimed
    conclusion.
        - `INCONCLUSIVE`: You cannot determine whether the proof is
    valid and cannot identify a decisive defect or a rigorous
    counterexample. Use this verdict when resolution depends on a
    result, computation, definition, or fact that cannot be verified
    from the supplied material or reliable standard knowledge. Do not
    use `INCONCLUSIVE` merely because the proof is difficult.

        After the verdict, provide the following labeled fields:

        STATEMENT:
        Restate the exact claim being audited, preserving its
    quantifiers and hypotheses.
```

```
    STRATEGY:
    Briefly summarize the proposed proof strategy without
improving it.

    CHECKS:
    List the decisive steps you independently checked, including
the hypotheses of each external result used.

    ANALYSIS:
    Give concise mathematical reasoning supporting the verdict.
For an invalid proof, identify the earliest decisive failure as
precisely as possible and explain why later steps do not repair
it. For a verified proof, give the decisive chain of implications.

    ROUTINE_COMPLETIONS:
    List any minor details you supplied, or write "None."

    DEPENDENCIES:
    List external theorems, citations, computations, or heuristic
assumptions on which the audit depends. Distinguish facts actually
verified from facts merely claimed in the supplied material.

    LITERATURE_RELATION:
    State whether the conjecture was found to be settled in a
revision or later work and whether the submitted proof reproduces,
invokes, differs from, or is
    independent of that proof. Write "None found" if no such
result was located.
    This field does not determine the proof-verification verdict.

    Use LaTeX notation. A separate pass will produce the final
compilable
    document.
```