First Proof Third Batch

October 14, 2026

Mohammed Abouzaid[1]
Stanford University

Nikhil Srivastava[2]
University of California, Berkeley

Rachel Ward[3]
University of Texas at Austin

Lauren Williams[4]
Harvard University

Problem Contributors

1. Leo Herr, *Virginia Tech*[5]; Jonathan Wise, *University of Colorado Boulder*[6]
2. Markus Reiß, *Humboldt-Universität zu Berlin*[7]; Sven Wang, *EPFL*[8]
3. Adam S. Sikora, *University at Buffalo*[9]
4. Bernd Siebert, *The University of Texas at Austin*[10]
5. Tuong Le, *Princeton University*[11]; Chayim Lowen, *Princeton University*[12]
6. Cameron Gordon, *The University of Texas at Austin*[13]
7. Tanya Firsova, *Kansas State University*[14]; Mikhail Mazin, *Kansas State University*[15]
8. Simion Filip, *University of Chicago*[16]
9. Benjamin Dozier, *Cornell University*[17]; Sidhanth Raman, *University of Chicago*[18]; Carlos Serván, *University of Michigan*[19]
10. Charles Doran, *University of Alberta*[20]; James M. Jones, *Philipps-Universität Marburg*[21]; Eric Pichon-Pharabod, *University of Oxford*[22]; Alan Thompson, *Loughborough University*[23].

Abstract

To assess the ability of current AI systems to correctly solve research-level mathematics problems, we tested several AI systems on our "third batch" of ten problems in a broad range of mathematical fields; these problems arose naturally in the research process of the contributors. This document includes the problems, our methodology, and the results of our testing. We assessed the AI-generated solutions based on mathematical correctness, quality of exposition, and quality of attribution and citations. We provide links to supplementary documents including the human solutions, the AI-generated solutions, and the referee reports and logs for the AI-generated solutions.

Contents

1 Introduction **3**
- 1.1 What is it that mathematicians accomplish? . . . 4
- 1.2 Background on First Proof . . . 5
- 1.3 Summary of Methodology . . . 5
- 1.4 What comes next . . . 6

2 Methodology **7**
- 2.1 Problem Selection (August–October, 2026) . . . 7
- 2.2 Testing (October 2–5, 2026) . . . 8
- 2.3 Reviewing (October 8–9, 2026) . . . 10
- 2.4 Funding . . . 12

3 Problems **13**
- 3.1 Problem 1: Herr and Wise (logarithmic geometry) . . . 13
- 3.2 Problem 2: Reiß and Wang (statistical inverse problems) . . . 14
- 3.3 Problem 3: Sikora (quantum topology) . . . 15
- 3.4 Problem 4: Siebert (symplectic geometry) . . . 17
- 3.5 Problem 5: Le and Lowen (hyperfields) . . . 17
- 3.6 Problem 6: Gordon (low-dimensional topology) . . . 18
- 3.7 Problem 7: Firsova and Mazin (complex dynamics) . . . 19
- 3.8 Problem 8: Filip (dynamical systems) . . . 20
- 3.9 Problem 9: Dozier, Raman, and Serván (Teichmüller theory) . . . 21
- 3.10 Problem 10: Doran, Jones, Pichon-Pharabod, and Thompson (algebraic geometry) . 22

4 AI Systems Tested **23**
- 4.1 System A: OpenAI GPT-6 Astra . . . 23
- 4.2 System B: Anthropic Siege (workflow on Claude Opus 5.5) . . . 24
- 4.3 System C: IMProofBench ProofCouncil . . . 24
- 4.4 System D: MerLean . . . 24
- 4.5 System E: Humanize (Humanfia) . . . 25

5 Results and Costs **25**
- 5.1 Problem 1 . . . 27
- 5.2 Problem 2 . . . 29
- 5.3 Problem 3 . . . 30
- 5.4 Problem 4 . . . 31
- 5.5 Problem 5 . . . 32
- 5.6 Problem 6 . . . 33
- 5.7 Problem 7 . . . 34
- 5.8 Problem 8 . . . 35
- 5.9 Problem 9 . . . 36

5.10 Problem 10 37

6 Open-weight submission: Humanize **38**
6.1 Referee assessments 39
6.2 Comparison with an older closed model 41
6.3 Further open-weight runs 42

7 Conclusion and takeaways **42**

8 Supplementary documents **44**

1 Introduction

The goal of our ongoing First Proof project is to provide accurate information to mathematicians and to the public about the capabilities of AI systems in the context of mathematical research. In light of the spectacular recent advances demonstrating the ability of AI systems to solve mathematical problems [4, 9, 7, 11], nearly all the recent media around AI and mathematics focuses on success stories. This makes it hard to get a nuanced, objective view of the current state of the public models, and indeed it runs the risk of making a single point of view all too pervasive.

From August–October 2026, we created and tested *Batch 3*, a collection of 10 problems designed to test the abilities of AI systems to provide proofs of specified mathematical statements. These 10 problems arose naturally in the research of our problem contributors, who had found solutions of these problems (of 12 pages or less), but had not yet publicized those solutions. We then tested five AI systems on these problems: in particular, we tested OpenAI's GPT-6 Astra; Anthropic's *Siege* (a workflow on Claude Opus 5.5); ProofCouncil, a multi-agent harness that calls multiple frontier models from OpenAI, Anthropic, and Google; MerLean, a harness that calls GPT-6 Astra and GPT-6.1 Sol; and Humanize, a harness that calls the **open-weight** model Kimi K3. After producing the AI-generated solutions to our problems, we formally graded the solutions, with the help of 30 expert mathematicians – and scored them based on **correctness**, **exposition**, and **attribution**. Recognizing that incorrect or poorly written solutions can waste significant time for the humans reading them, our rubric awarded scores to each AI-generated solution from -1 (worst) to 1 (best). Our process was overseen by the *First Proof Foundation*.

Batch 3 came on the heels of *Batch 2* [2], a similar collection of 10 problems that we created and tested from March–June 2026. In broad strokes, the results of Batch 3, with comparisons to Batch 2, are as follows. (More details can be found in Section 7.)

Frontier systems have improved. The ProofCouncil harness and GPT-6 Astra each answered two more problems correctly than in the previous batch, obtaining 9 out of 10 and 7 out of 10 respectively. ProofCouncil and GPT-6 Astra respectively cost $1,888 and $36 (compared to $3,200 and $117 in Batch 2). Each of the 10 questions in Batch 3 was solved correctly by some system, setting aside citation and exposition issues.

Open-weight systems are far behind. Our results indicate that when it comes to research mathematics, the gap between open-weight models and closed systems is substantially wider than the roughly four-month lag that Epoch AI measured across general AI benchmarks in May 2026. ProofCouncil solved 9 of the 10 problems, while the system built on the open-weight model Kimi K3 solved none. In particular, every submission by the open-weight system contained major errors, at a total cost of about $2,000. Reviewers compared it to older frontier models and found it generally less capable than the leading frontier model of December 2025 (OpenAI's GPT-5.2 Pro).

There is a substantial cost for improved performance. By leveraging more compute, it is possible to significantly improve on the performance of a single query to GPT-6 Astra – but this comes at significant financial cost, leading to concerns that access to computational resources may become a barrier for basic research in mathematics.

Exposition and attribution still lag. Frontier models cite the literature substantially better than they did in Batch 2, but gaps remain. Only 4 of the 10 GPT-6 Astra solutions were judged to have adequate citations; the ProofCouncil harness, which uses Astra as a base model, was able to improve this to 7 out of 10. Exposition remains generally weak: even the top system ProofCouncil had a third of its responses receive negative average scores.

In the remainder of this introduction, we explain more about how we think about mathematics, and provide more details on our methodology.

1.1 What is it that mathematicians accomplish?

This is the title of the first section of William Thurston's famous 1994 essay "On proof and progress in mathematics" [15]. He writes:

> There are many issues buried in this question . . . it would not be good to start, for example, with the question 'How do mathematicians prove theorems?' . . .
>
> The question is not even 'How do mathematicians make progress in mathematics?' Rather, as a more explicit (and leading) form of the question, I prefer 'How do mathematicians advance human understanding of mathematics?'
>
> This question brings to the fore something that is fundamental and pervasive: that what we are doing is finding ways for *people* to understand and think about mathematics.

There is a misconception in the popular imagination that there is a finite list of 'holy grail' mathematics conjectures, and that professional mathematicians make their living by trying to make their way to one or another of them. However, much of our work is to push the boundary of mathematics by asking new questions and developing frameworks to answer these questions. This gives novel ways for people to understand and think about mathematics.

At present we do not have concrete proposals for devising experiments that would measure the ability of AI systems to produce interesting new conjectures, or useful mathematical definitions and frameworks. In this third batch, we have thus, as in the first two batches, restricted our attention to

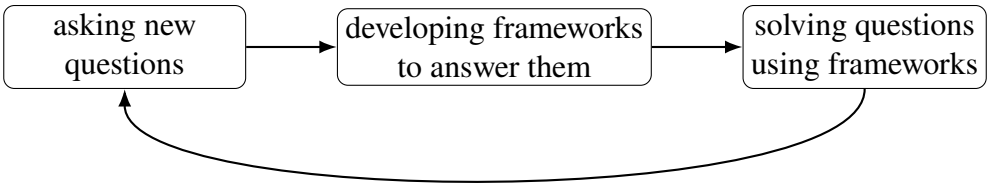


Figure 1: The mathematical research cycle.

assessing mathematical proofs that AI systems have created. As before, we collected solved but unpublished mathematical problems from human mathematicians, which we posed as a challenge to AI systems. But this time, we tried to highlight the goal of "finding ways for *people* to understand and think about mathematics," by evaluating solutions not only on correctness, but on quality of exposition and attribution.

1.2 Background on First Proof

The first two batches of problems for First Proof were collections of 10 problems designed to test the abilities of large language models to provide proofs of specified mathematical statements. We released our first batch of 10 problems in February 2026 [1] as an informal collaborative experiment, followed by a second batch of 10 problems in May 2026 [2] which was designed as a formal benchmark. Each set of problems consisted of statements which had come up naturally in the research of mathematicians – but whose proofs had not appeared on the internet or been published. In Batch 1, we allowed anyone in the community to engage with the problems and we did not review the solutions; we instead used Batch 1 as an experiment which informed our design of the second batch. In Batch 2, we invited commercial companies and academic teams to participate in our benchmark, and, guided by our goal of transparency (*cf.* Section 2) we ultimately tested four systems. We then brought together thirty expert mathematicians who spent two days reviewing the AI-generated solutions for mathematical correctness. While we saw some impressive successes among the solutions – a total of 7 problems were solved by the union of all systems considered – we also saw that there were certain issues that came up repeatedly in the referee reports. A main takeaway from Batch 2 was that even when an AI system produced a solution that was deemed to be mathematically correct, the solution was often very **hard to understand**, or **failed to properly cite the sources** it was building on. This takeaway is what motivated us to update our scoring system for Batch 3.

1.3 Summary of Methodology

In keeping with the values of the academic math community, our guiding principle in designing the second and third batches was **transparency**. We decided that in order to conduct an experiment in which we could have total clarity about prompts and human interaction, we needed to run all of the tests ourselves. Since our goal is to inform mathematicians and the public about models that

are in principle available to them[1] we decided to test only those AI systems that were public at the time of testing, or to test harnesses that call only publicly available models and would be released to the public. We also committed to releasing a full account of our process: the ten problems, the human-generated solutions, the AI-generated solutions, the cost of creating those solutions, our evaluations of those solutions, the source code of the AI systems modulo publicly available models, and the logs produced by the AI systems.

This third batch consists of a benchmark whose results we report here. Our methodology for selecting, testing, and reviewing the 10 problems comprising our benchmark is described in detail in Section 2. A high-level summary indicating differences between the second and third batches is provided in Table 1. The most important differences are that in this third batch, we assessed[2] the AI-generated solutions not only on mathematical correctness, but also on quality of exposition and quality of attribution.

	Batch 2 Benchmark (March–June'26)	**Batch 3 Benchmark (August–October'26)**
Problem Selection		
Number of problems	10	10
Problem source	Research mathematicians	Research mathematicians
Max proof length	8 pages	12 pages
Preliminary AI testing	ChatGPT 5.5-high/xhigh, Gemini 3.1 Pro, Opus 4.7	ChatGPT 5.6 Sol Pro, ChatGPT 5.6-med/high/xhigh, Opus 5
Testing		
Who may participate	Open-source harnesses calling public models	Open-source harnesses calling public models
Who runs the tests	First Proof	First Proof
Reviewing		
Process	Double-blind review with ∼30 expert referees	Double-blind review with ∼30 expert referees
Review Rubric	Based on math journals: accept/minor/major revisions/reject	Separate scores for correctness, exposition, and attribution, with negative scoring for wrong solutions

Table 1: Similarities and differences between Batch 2 and Batch 3 Benchmarks.

1.4 What comes next

Going forward, we would like to put more emphasis on testing of open-weight models. We would also like to test more open-ended mathematical questions. Batch 4 of First Proof will take place from October 2026 to January 2027, with results released in the first half of January.

[1] Given sufficient funding.

[2] with the help of 30 expert mathematicians

If you are a mathematician interested in contributing to First Proof, please reach out to us at contact@1stproof.org.

The structure of this paper is as follows. In Section 2 we describe our methodology, including the problem selection, testing process, reviewing, and funding. We then give the statements of all ten problems in Section 3, along with comments by the authors of those problems on context and novelty. In Section 4 we describe the AI systems that we tested, and in Section 5 we summarize the results of the testing, including numerical scores as well as comments. One of the AI systems that we tested was an open-weight model; we treated this one separately from the rest, and discuss its results in Section 6. In Section 7 we summarize our conclusions and takeaways from Third Batch, and in Section 8 we provide links to supplementary documents.

AI Statement

Claude Opus and ChatGPT Astra were used to assist with writing some of the review summaries in Section 5 of this paper. As described in Section 2.3, reviewers were allowed to use LLMs to help them understand and find mistakes in the AI-generated solutions, but were asked to make judgments themselves.

Acknowledgements

We thank the Center of Mathematical Sciences and Applications (CMSA) at Harvard University for hosting the problem-writing session, and the Institute for Computer-Aided Reasoning in Mathematics (ICARM) at Carnegie Mellon for hosting the reviewing session for Batch 3. We are particularly grateful to the 30 referees who came to ICARM on short notice to evaluate the AI-generated solutions to Batch 3.

2 Methodology

2.1 Problem Selection (August–October, 2026)

Solicitation. We solicited problems from mathematicians representing a wide range of mathematical fields as well as geographic locations (mostly in the United States for logistical reasons). We aimed to obtain mathematical problems which involve a nonstandard insight to answer, and which have a proof known to the mathematician of **at most 12 pages** which had not appeared on the internet or in any public forum.

We asked that the authors:

- not discuss the candidate problems with any AI system before the problems were announced, except in a controlled zero data retention environment (see below).
- produce a **solution** of at most 12 pages which consists of a complete proof, with accurate references to the mathematical literature;

- pre-register a **descriptive statement**, written for a general audience of mathematicians, describing the (i) context of the problem, the (ii) novelty of the argument, (iii) the general strategy of the proof, and (iv) an estimate of the difficulty of the problem, e.g. via the time that it has taken to solve the problem.
- confirm that they will not consult for companies with a direct interest in this benchmark while working on this project.

The solicitation yielded **25 candidate problems**.

Preliminary AI Tests. We ran preliminary AI tests on each candidate problem, because we wanted to have high confidence about the following: (i) the problem is comprehensible to an LLM; (ii) there are no loopholes or forgotten assumptions that would render the problem trivial; (iii) it is not straightforward for the AI models available at the time of pretesting.

We ran the bulk of our preliminary tests during August 13–14, 2026, with a few additional tests over the following two weeks. In these tests, the authors of the problems probed properties (i)–(iii) non-interactively using a fixed one-shot prompt on ChatGPT 5.6 Sol Pro (xhigh) with internet disabled, and interactively using ChatGPT 5.6 Sol (medium-high) on OpenRouter, both with zero data retention. Additionally, over August 10–12, First Proof tested all problems using a fixed one-shot prompt on ChatGPT 5.6 Sol Pro (xhigh), with internet search disabled, directly via the cloud computing platform Azure. A few problems were retested during the final two weeks of August using the same specifications, to make sure the problems satisfied properties (i)–(iii).

Eight problems were discarded at this stage: four were solved or close to solved by AI in pre-testing, three were judged too easy, and one was judged by a human referee to be at the level of an olympiad problem. **Two further problems** were withdrawn by their authors.

Selection. From the remaining 15 problems, we selected **10 problems for the benchmark**, based on the following criteria: (i) balance across fields of mathematics; (ii) balance across geographic locations; (iii) balance across perceived difficulty and novelty by the authors; (iv) availability of referees on short notice.

Human Solution Refereeing. For each of the selected problems we asked the author to write a solution. We had all (human-generated) solutions go through a first round of refereeing by a single human expert referee — suggested by the author of the problem — to check clarity and correctness. Feedback from the referee reports was used to revise some of the solutions.

2.2 Testing (October 2–5, 2026)

Eligibility. We announced the third batch on August 25, 2026, stating that eligible systems were open-source harnesses that call publicly available models via API; a single prompt is a special case. We asked interested parties to contact us by September 10, 2026. Guided by our goal of transparency, we put the following constraints on AI systems to be tested in the third batch benchmark:

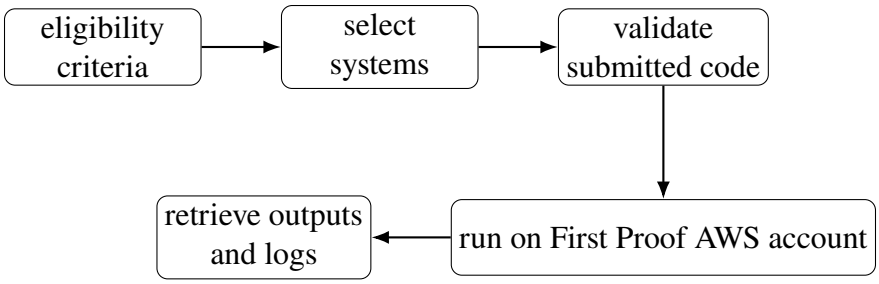


Figure 2: Testing process.

1. The system must consist of code (which we will sometimes refer to as a **harness**) which calls publicly available models via API. Here, *publicly available* means that any member of the public must be able to access the model as of October 7, 2026.

2. Source code, prompts, and API keys must be supplied by October 1, 2026.

3. The system must take as input ten math problems given as .tex source, solve them in one shot (i.e., with no further interaction), run for at most 24 hours, and produce ten Latex solution documents of at most 16 pages each.

4. The system must log input, output, and reasoning tokens. We did not put an upper bound on the number of tokens allowed.

5. First Proof must be allowed to publish all of the output, logs, and code on October 14, 2026.

Note that the above constraints allow as a special case a harness which just calls a public model using a one-shot prompt for each problem.

Selection of Systems to Test. Based on the above criteria and discussions with interested parties, it was decided in late September that we would test five AI systems: GPT-6 Astra from OpenAI, a Claude Opus 5.5 skill from Anthropic, and three harnesses from academic teams. These are described in detail in Section 4.

Testing Implementation. We required that each participating system be implemented as a complete, reproducible pipeline that could be deployed on standard cloud infrastructure.

All systems processed a common JSON input file containing the 10 benchmark problems. *For each submission except for the open-weight submission (Humanfia)*, the testing implementation was essentially identical to the Batch 2 specifications, with First Proof owning the infrastructure. For each system, First Proof launched a fresh ephemeral CPU-only EC2 instance with 8 vCPUs, 64 GB of RAM, and 100 GB of disk storage. First Proof copied over the code and input file, built a Docker container, and executed it. Containers had outbound internet access for API calls and limited fetching like paper PDFs; inbound network access was restricted via security group to the operator's IP. The container read the input file and wrote its results to a designated output directory. Submitters provided any required API keys through a secure out-of-band channel. After the container exited

or hit a 24-hour timeout, First Proof retrieved the outputs and logs, recorded run metadata, and terminated the instance.

The Humanfia harness ran under the same protocol, but the Kimi K3 model it called was not accessed through a public API provider. Instead, Humanfia self-hosted the open-weight Kimi K3 model [8], using the official weights [10], on two servers with eight NVIDIA B300 GPUs each, provided through DeepInfra. Each server ran a full copy of the model and used speculative decoding with the Kimi-K3-DSpark draft model [13], which speeds up generation without changing the distribution of the model's outputs.

2.3 Reviewing (October 8–9, 2026)

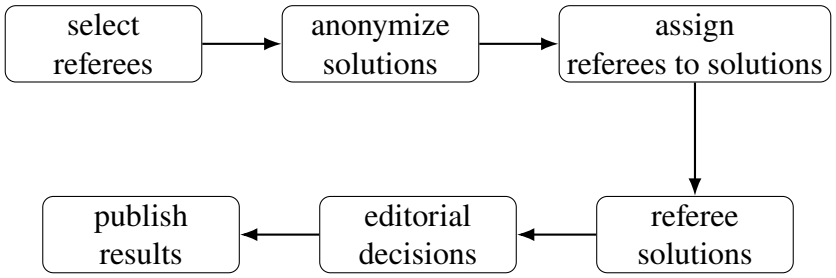


Figure 3: Review process.

Selection of Referees. We asked the authors of the problems to suggest a list of referees for each problem, that is, mathematicians who had expertise in the field and were qualified to review the solutions. After several rounds of invitations, we were able to confirm 3 referees for each problem. The total number of confirmed referees across all problems was 30.

Anonymization and Referee Assignment The solutions produced by the AI systems were given randomized animal names (Puffin, Kestrel, Marten, etc.), with the identities of the corresponding AI systems visible only to the editors. Each submission was assigned 2 or 3 referees based on its perceived intricacy by the editors, which in turn was partially based on quick opinions from the problem authors.

Refereeing Process. The anonymized solutions produced by the AI systems were reviewed by the human referees in a manner similar to the review process in mathematics journals, during an in-person meeting at Carnegie Mellon's Institute for Computer-Aided Reasoning in Mathematics (ICARM) during October 8–9, 2026.

We asked the referees to give three scores to each solution based on the **correctness of the solution**, the **quality of exposition**, and **quality of attribution** and **citations**. We also encouraged the referees to comment on the novelty of the solution, i.e. whether it contained new ideas versus standard techniques.

Each submission was assigned three referees, except Humanize on Problem 10, which was assigned two. For 41 of the 45 submissions, at least two referees gave a correctness score; for Humanize on Problems 2 and 7 and for Siege on Problem 7 only one did, and no referee could determine correctness for Siege on Problem 6. Each solution received both qualitative and quantitative feedback, using the following scales. When in doubt between choosing one of two scores, referees were asked to make a judgment and choose one of them.

Correctness.

- 1: essentially flawless mathematical argument
- 0.5: minor errors that are easily corrected
- 0: blank solution, or acknowledged partial result
- -0.5: major errors that require substantial mathematical work to correct
- -1: no substantial progress / unsalvageable approach

Exposition.

- 1: well-written and clear solution.
- 0: argument that is possible to follow with serious effort
- -1: incomprehensible

Attribution.

- 1: essentially appropriate attributions to related work and theorems being used
- 0: moderate issues with missing attributions or hallucinated references
- -1: blatant plagiarism or hallucinations

We decided to use negative numbers in our scale because spending time reading solutions that are incorrect, poorly written, or poorly attributed, can be a substantial waste of time for a human. In particular, on the correctness axis, we wanted to incentivize systems to output no answer or admit they couldn't solve the problem, rather than to give a deceptive wrong solution.

Referees were instructed to spend the first three hours of the meeting at ICARM reading the submissions individually. After this initial reading, referees were allowed to ask each other mathematical questions for clarification, *were allowed to use LLMs* to aid their reading and to search for errors and gaps, and were instructed to make their own judgments about assessments.

Editorial Decisions. The Editorial Board interviewed the referees, then summarized the referee reports taking the conversations into account. These summaries along with the scores are given in Section 5.

Publication. All problems, author solutions, author pre-registered statements, tested systems' source code, AI solutions, referee reports, and editorial decisions, along with editorial reasoning, were published on October 14, 2026.

2.4 Funding

Batch 3 is supported in part by charitable foundations, the Survival and Flourishing Fund, the AI for Math Fund of Renaissance Philanthropy, and Google.org, so that we can test systems from non-commercial entities. It is also supported in part by unrestricted donations from Anthropic and OpenAI, used exclusively for scientific and outreach activities; no member of either board is paid from these funds. This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under the Exponentiating Mathematics (expMath) program, Agreement No. HR0011263E126. The views, opinions, and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

The First Proof Foundation is registered as a 501(c)(3) corporation and it releases full reports of our expenses.

3 Problems

3.1 Problem 1: Leo Herr and Jonathan Wise (logarithmic geometry)

In the following, a monoid is an integral, saturated, sharp, commutative, unital semigroup. Monoids are written additively. A monoid M induces a translation-invariant partial order on its associated group M^{gp} such that $x \geq 0$ if and only if $x \in M$. A valuative monoid is a monoid in which the associated partial order is a total order.

If M is a monoid, $x \in M^{\mathrm{gp}}$, and $y \in M$, we write $x \prec y$ to mean that there are integers m and n such that $my \leq x \leq ny$.

A monoid homomorphism $f : M \to N$ is called *infinitesimal* if the following condition holds:

- If there are $x \in M^{\mathrm{gp}}$ and $y \in M$ such that $f(x) \in N$ and $x \prec y$ and $f(y) \prec f(x)$ then $x \in M$.

If $P \xrightarrow{f} Q \xrightarrow{g} N$ are monoid homomorphisms, we say that f is *birhizomic* at g if it is of finite presentation in the category of (integral, saturated, sharp) monoids and, for every factorization of gf through monoid homomorphisms $P \xrightarrow{h} M \xrightarrow{p} N$ with p infinitesimal, there is a unique monoid homomorphism $u : Q \to M$ with $pu = g$ and $uf = h$.

A family of morphisms $P \to Q_i$ is called a birhizomic cover if, for every valuative monoid V and every homomorphism $v : P \to V$, there is a factorization of v through some $Q_i \to V$ such that $P \to Q_i$ is birhizomic at $Q_i \to V$.

Let P be a monoid. Let A_P be the category of monoid homomorphisms $P \to Q$. The birhizomic covers form a Grothendieck topology on A_P^{op}.

Let P be a monoid. For any x in P, there is an initial monoid homomorphism $P \to P[-x]^\sharp$ that sends x to 0. Let B_P be the full subcategory of A_P consisting of homomorphisms $P \to P[-x]^\sharp$, for $x \in P$.

We have given A_P^{op} the birhizomic topology and B_P^{op} the trivial topology in which all presheaves are sheaves. The inclusion $B_P^{\mathrm{op}} \subset A_P^{\mathrm{op}}$ is left exact. By left Kan extension and sheafification, it extends to a functor $\pi^{-1} : \mathrm{Sh}(B_P) \to \mathrm{Sh}(A_P)$ on the categories of sheaves. Let π_* be the right adjoint of π^{-1}.

Is it the case that, for all presheaves of abelian groups F on B_P^{op}, the morphism

$$F \to \mathrm{R}\pi_*\pi^{-1}F$$

is a quasi-isomorphism?

Pre-registered statement.

Logarithmic geometry is an amalgam of algebraic geometry and piecewise linear, or tropical, geometry. This problem is drawn from a larger project concerning topologies underlying logarithmic geometry. We are looking for a topology similar to the logarithmic étale topology but with better formal properties. One such candidate is the 'rhizomic topology' mentioned in the problem statement. This problem is part of the demonstration that many sheaves in the rhizomic topology have better descent properties than they do in the logarithmic étale topology.

In its broad structure, our proof follows a familiar outline: first, we make a finite type approximation to reduce the question to monoids of finite type; next, we reduce to a smaller class of sheaves by the choice of a suitable resolution; finally, we calculate the cohomology of those sheaves using an explicit simplicial complex. It took us some time to identify the correct approach in the second step, but the answer is ultimately one of several standard resolutions. The first and third steps required what we felt were novel arguments occasioned by the setting. We drew inspiration from noetherian approximation, the cotangent complex, and the étale topology in algebraic geometry. However, the proofs of analogous statements required arguments that were (to our eyes) completely different, and took us some effort to discover.

The principal difficulties are likely 1) identifying the appropriate analogy to pursue, and 2) implementing the analogy in a novel setting, where there is relatively little relevant literature. While we feel that the geometric solutions we found in the second stage might not be reproduced by an AI without prompting, we would not be surprised if an AI could find another approach.

3.2 Problem 2: Markus Reiß and Sven Wang (statistical inverse problems)

Let $D \subseteq \mathbb{R}^d$, $d \geq 2$, denote the d-dimensional Euclidean unit ball. Given a function $f \in C^\alpha(D)$ with Hölder-Zygmund smoothness $\alpha > 1$ on D and let $u = u_f$ denote the unique solution of the following divergence form PDE (Darcy's problem)

$$\mathrm{div}(f\nabla u) = 1 \ \text{ on } D, \quad u|_{\partial D} = 0.$$

Consider the non-linear statistical inverse problem of reconstructing f from noisy measurements of u corrupted by Gaussian white noise

$$Y_n = u_f + \frac{1}{\sqrt{n}}\mathbb{W}.$$

For $r \in (0,1)$ consider the following class of conductivities

$$f \in \mathcal{F}(\alpha, r) = \{h \in C^\alpha(D),\ \|h - \mathbf{1}\|_{C^\alpha(D)} \leq r\},$$

where $\mathbf{1}$ denotes a constant function with value 1. For small (but fixed) $r \in (0,1)$, what is the minimax rate as $n \to \infty$ for estimating f in the L^2-mean squared error

$$\inf_{\hat{f}_N} \sup_{f_0 \in \mathcal{F}(\alpha,r)} \mathbb{E}_{f_0}\Big[\|\hat{f}_n - f_0\|^2_{L^2(D)}\Big],$$

up to logarithmic factors?

Pre-registered statement.

Problem context & motivation. The Darcy problem is a prototypical non-linear inverse problem where the goal is to estimate diffusion coefficient of an elliptic PDE [14, 12, 5]. In the multi-dimensional ($d \geq 2$) case, this model has been a central example in the methodological literature

on non-linear inverse problems, including for infinite-dimensional Markov Chain Monte Carlo (MCMC) methods. The problem of determining the statistical *minimax-optimal rates* is a natural statistical question in this context, and has been discussed at conferences on mathematical statistics in the past years. Different convergence rates have been derived in the last 10 years, however, the minimax-optimal convergence rate is unknown in any dimension. Here, we pose the question for $d = 1$, a simpler case where the forward map is explicitly given.

Strategy of human proof. For $d = 1$ there is an explicit characterisation of the forward map $f \mapsto u$ from which one can read off the 'mixed ill-posedness' of the problem. This allows to perform local bias-variance computations for kernel-type estimators with locally varying bandwidth, which take into account the distance to the unique root x_0 of u'. Due to this spatial inhomogeneity, the bias calculation is non-standard and requires careful analysis. Moreover, the root x_0 of u' is unknown; thus one employs sample splitting to obtain an independent pilot estimate of x_0 and studies the plug-in estimator arising from usage of the estimated value of x_0. Integrating the local pointwise risk with two regimes leads to an "elbow" at $\alpha = 3/2$. The lower bounds arguments are more classical, given the upper bound analysis, but the alternatives in the multiple hypothesis testing problem must have certain cancellation properties to ensure x_0 remains fixed and must again be chosen with a suitable locally varying scaling.

Novelty. Due to the mixed ill-posedness of the problem, standard approaches borrowed from linear inverse problems or arising from 'forward smoothing' properties together with 'inverse stability' estimates between Sobolev-type norms fail. Thus, no result is currently available characterizing the optimal rate of convergence (in either L2 or pointwise loss), in either the deterministic setting and in the present statistical setting. The convergence rate with 'elbow' at $\alpha = 3/2$ is atypical and non-standard in non-parametric statistics.

Difficulty. MR and SW had about 10 in-depth blackboard discussion on-and-off over the course of one year, as well as at least 5 full working days per person. We had discussions with several leading experts in the field and there was no uniform opinion on what the minimax-optimal convergence rate should be. The multi-dimensional case can be expected to be considerably harder than the case $d = 1$.

3.3 Problem 3: Adam Sikora (quantum topology)

Let $\overline{\mathscr{S}}(\mathfrak{B}^*)$ be the reduced $\mathrm{SL}(n)$-stated skein algebra of the punctured bigon, $\mathfrak{B}^*$, considered as an algebra over the field $\mathbb{K} = \mathbb{Q}(v)$, where $v = q^{\frac{1}{2n}}$ for some fixed $n \geq 2$. (It is the stated skein algebra of Lê–Sikora, quotiented by the bad-corner-arc ideal of Lê-Yu.)

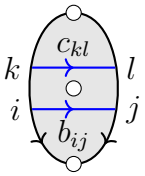
Let $\check{\mathcal{U}}_q(\mathfrak{b}_+), \check{\mathcal{U}}_q(\mathfrak{b}_-)$ be the upper and lower Borel subalgebras of the simply-connected quantum group $\check{\mathcal{U}}_q(\mathfrak{sl}(n))$. (This quantum group appears for example in the Klimyk–Schmüdgen book as $U_q^{ext}(\mathfrak{sl}(n))$.)

(1) Construct an explicit algebra isomorphism between $\overline{\mathscr{S}}(\mathfrak{B}^*)$ and the generalized quantum double $\check{\mathcal{U}}_q(\mathfrak{b}_-) \bowtie \check{\mathcal{U}}_q(\mathfrak{b}_+)$ (with respect to the standard Drinfeld–Tanisaki skew pairing,

$$\langle \hat{K}_i, \hat{K}_j \rangle = q^{\frac{1}{n} - \delta_{i,j}}, \quad \langle E_i, F_j \rangle = \delta_{i,j}/(q^{-1} - q), \quad \langle \hat{K}_i, F_j \rangle = \langle E_j, \hat{K}_i \rangle = 0,$$

in the notation of Klimyk–Schmüdgen).

(2) Deduce that the algebra $\check{\mathcal{U}}_q(\mathfrak{sl}(n))$ is isomorphic with $\overline{\mathscr{S}}(\mathfrak{B}^*)/I$, where I is the two-sided ideal generated by $b_{ii}c_{ii} - 1$ for $1 \le i \le n$, and b_{ij}, c_{kl} are the horizontal arcs directed to the right:



Pre-registered statement.

Problem context. Quantum Groups play a central role in algebra, representation theory and low-dimensional topology over the last four decades. Drinfeld's quantum group is (essentially) the quantum double of positive and negative Borel parts. The product in this construction is rather abstract and complicated. The problem asks for a "graphical" realization of this quantum double for $\mathfrak{g} = \mathfrak{sl}(n)$, as an algebra built of certain graphs (called "webs") subject to certain local "skein" relations. (Such construction is called a skein algebra.)

In our problem we ask about an algebra isomorphism only. However it is easy to define a bialgebra structure on $\bar{S}_n(\mathfrak{B}^*)$, and on $\bar{S}_n(\mathfrak{B}^*)/I$ (double the puncture and split $\mathfrak{B}^*$ into $\mathfrak{B}^* \cup \mathfrak{B}^*$) and to show that this is the coproduct on $\check{\mathcal{U}}_q(\mathfrak{sl}(n))$.

This problem comes from a joint work with Thang T. Q. Lê.

Novelty. For $n = 2$, the result was proved by Korinman. Higher rank is a non-trivial generalization since webs have vertices and additional local relations, and the interaction of webs on opposite sides of the puncture is substantially more complicated. The new insight is that these local topological interactions encode precisely the otherwise abstract mixed relations of the quantum double, with no additional relations needed.

Difficulty. The proof requires expertise in quantum groups, Hopf pairings, and stated skein algebras, together with geometric arguments. We regard it as a substantial research-level problem rather than an application of existing results.

Strategy of the proof. The proof requires observing isomorphisms

$$\check{\mathcal{U}}_q(\mathfrak{b}_-) \simeq \mathcal{O}_q(B_+), \quad \check{\mathcal{U}}_q(\mathfrak{b}_+) \simeq \mathcal{O}_q(B_-)$$

and realizing that $\mathcal{O}_q(B_+)$ and $\mathcal{O}_q(B_-)$ correspond to the lower and upper parts of the reduced skein algebra of a punctured bigon. Now the key step is to show that moving upper webs past the lower ones gives exactly the mixed relations determined by the Hopf pairing in the quantum double. Thus, the puncture geometrically produces the twisting in $\check{\mathcal{U}}_q(\mathfrak{b}_-) \bowtie_\rho \check{\mathcal{U}}_q(\mathfrak{b}_+)$.

3.4 Problem 4: Bernd Siebert (symplectic geometry)

Let (M, ω) be a closed symplectic manifold and $D \subset M$ a symplectic simple normal crossings divisor. Does there exist an ω-tame almost complex structure J such that locally on M there are smooth complex coordinates $(z_1, \ldots, z_n)$ in which, for some $0 \leq r \leq n$, the functions $z_1, \ldots, z_r$ are J-holomorphic and D is given by

$$z_1 \cdots z_r = 0?$$

Here one may first replace ω by a symplectic form connected to it through a path of symplectic forms, all cohomologous to ω, along which D remains a symplectic simple normal crossings divisor.

Pre-registered statement.

The statement is a strengthening of a result by Tehrani/McLean/Zinger from 2014. That result constructs an almost complex structure making each component of the snc divisor $D = D_1 \cup \ldots D_r$ an almost complex submanifold. The requested almost complex structure is a lot more restricted by asking for the existence of enough pseudo-holomorphic functions locally to cut out the components of D. Note that a general almost complex structure has no non-constant pseudo-holomorphic functions, and that if D has a zero-dimensional stratum then the existence of locally defining functions implies that the almost complex structure is integrable.

The theorem is a technical result needed to easily import many of the techniques developed in the context of Gromov-Witten theory for the pair (X, D) developed in algebraic geometry to the symplectic setting. It might also simplify other questions concerning the symplectic geometry of a pair (X, D).

A first observation is that if such functions exist, then the normal bundles $N_{D_i|X}$ have a system of pseudo-holomorphic transition functions. Conversely, if one has an appropriately compatible system of pseudo-holomorphic structures on the normal bundles $N_{D_i|X}$, the mentioned TMZ regularization result almost directly implies the existence of the requested almost complex structure. Another rather obvious idea is that as in complex geometry, one can produce such pseudo-holomorphic line bundles by pseudo-holomorphic divisors in analogy to $\mathcal{O}_X(D)$. It is then natural to import a standard argument in algebraic geometry to twist with a very positive hypersurface to realize arbitrary complex line bundles in the form $\mathcal{O}_X(D_1 - D_2)$ for two effective divisors D. This argument requires a rather straightforward generalization of Donaldson's existence theorem for very positive symplectic hypersurfaces with the additional requirement of transversality with respect to D.

The subtle part of the proof that took me several weeks to get right is to arrange the auxiliary hypersurface and choice of representatives $D_1 - D_2$ of the normal bundles on each stratum of D in a compatible fashion.

3.5 Problem 5: Tuong Le and Chayim Lowen (hyperfields)

Let $\mathbb{H}$ be a hyperfield with multivalued addition operator $\boxplus$. Let $x, y, z \in \mathbb{H}$ be elements. Suppose that $0 \in x \boxplus y \boxplus z$. Does it then follow that

$$0 \in x^3 \boxplus y^3 \boxplus z^3 \boxplus -xyz \boxplus -xyz \boxplus -xyz?$$

Pre-registered statement.

Problem context. Hyperfields were introduced by Krasner in the 1950s in order to be able to compare valued fields of *different* characteristics. They would receive scarce attention until the turn of the century. A notable exception was Massouros's construction in the 1980s of a hyperfield which does not arise from a field using Krasner's construction. Connes and Consani, beginning in the late 2000s, have advanced a program to study the adèle class space using this framework. Around the same time, Viro showed that hyperfields provide a natural language for tropical geometry. More recently, Baker and Bowler showed that hyperfields can be used to unite and generalize the notions of valuated, oriented, and ordinary matroid using the framework of realizations over hyperfields, naturally extending linear realizability over fields.

These developments have sparked renewed interest in these field-like objects. In their work, Baker and Bowler were naturally led to introduce a much less rigid structure called a *tract*. Every hyperfield can naturally be interpreted as a tract, and so can the *subobjects* of hyperfields. These should be to hyperfields what integral domains are to fields. The absence of zero-divisors is a straightforward condition on commutative rings which precisely characterizes those that embed into a field. Famously, there is no similarly simple characterization of the subrings of skew fields. Likewise, we do not have a good characterization of the subtracts of hyperfields. If this question is to have a reasonable answer, it will necessarily give a characterization of the implications between algebraic equations in a hyperfield. Our question is the simplest possible non-trivial test case for such an implication.

Novelty. Our proof is a "just-do-it" construction reminiscent of the construction of the algebraic closure of a field by the procedure which may cheekily be called "just keep adding roots". Our non-explicit method of constructing a hyperfield appears to be new. The main insight needed is to have the correct framework. The intuition that tracts are the natural ambient category in which to work and that hyperfields ought to be considered as special members is a rather recent insight due to Baker–Bowler, Baker–Lorscheid and others. To even get started, one crucially needs a characterization of hyperfields among tracts. A particularly beautiful such characterization was given by Pedro Souza. From here, the proof is a matter of careful bookkeeping and the important observation that the expression $x^3 + y^3 + z^3 - xyz - xyz - xyz$ is *short enough*.

Difficulty. There are a number of ways to construct hyperfields. For instance, one can use Krasner's construction starting with a field or Massouros's construction and its variants. In recent work, we showed how to almost surely construct a finite hyperfield using a probabilistic method. The surprising conclusion we reached after examining the constructions above is that none can be used to produce a negative answer to our question. The appendix to our solution is dedicated to an explanation of this difficulty. We do not see an obvious approach that sidesteps the (implicit or explicit) use of tracts.

3.6 Problem 6: Cameron Gordon (low-dimensional topology)

Does a 3-manifold obtained by Dehn surgery on an L-space knot have at most two prime summands?

Pre-registered statement.

The context of this question is Dehn surgery on knots. This is a construction for producing 3-manifolds from knots by removing from the 3-sphere a neighborhood of a knot K and sewing it back differently. The resewing is parametrized by a rational number r, so one denotes the resulting 3-manifold by K(r). The main open problem in the area asks when K(r) can be a non-trivial connected sum: the Cabling Conjecture (CC) of González-Acuña and Short (1986) asserts that this only happens in the very special case that K is a cable knot (and r is the cabling slope). A consequence of the CC would be the Two Summands Conjecture (2SC) (Howie, 2010) that K(r) never has more than two prime summands. This is also open.

A couple of years ago I was trying to prove the CC for L-space knots, a class of knots that arises naturally in the context of Heegaard Floer homology. This is still an ongoing project, but at least I came up with a simple argument that gave some information, including a proof of the 2SC for such knots. This is the question posed to AI.

The proof is short and not particularly difficult. It makes use of several published results, in particular the main result of Hom, Lidman and Zufelt, "Reducible surgeries and Heegaard Floer homology", Math. Res. Lett. 22 (2015), 763–788, and the known structure of any K(r) that has more than two prime summands (see Corollary 5.3 of Howie, "A proof of the Scott–Wiegold conjecture on free products of cyclic groups", J. Pure Appl. Algebra 173 (2002), 167–176). The difficulty is not so much in the proof itself as in deciding how to approach the problem. It's hard for me to say how long it took as I was trying various approaches (some still ongoing) and of course I was really interested in the CC rather than just the 2SC (for L-space knots). The approach gives a bit more than the 2SC but not the full CC.

3.7 Problem 7: Tanya Firsova and Mikhail Mazin (complex dynamics)

Let $f : \mathbb{S}^2 \to \mathbb{S}^2$ be a branched cover, $V(f)$ be the set of its critical values, and $A, B \subset \mathbb{S}^2$ be finite sets.

1. If $f(A) \cup V(f) \subset B$, then every complex structure on $(\mathbb{S}^2, B)$ can be pulled back to a complex structure on $(\mathbb{S}^2, A)$, and this induces the map $f^* : \mathrm{Teich}(B) \to \mathrm{Teich}(A)$.

2. If $A \subset B$, then forgetting the points $B \backslash A$ induces the map $i_* : \mathrm{Teich}(B) \to \mathrm{Teich}(A)$.

If $f(A) \cup V(f) \cup A \subset B$, then the deformation space is defined by

$$\mathcal{D}_f := \{\tau \in \mathrm{Teich}(B) : f^*(\tau) = i_*(\tau)\}.$$

Consider a quadratic rational map f with critical points c_1 and c_2. Assume that c_1 belongs to a cycle of length 4, and the critical value $v = f(c_2)$ does not belong to the cycle. By taking

$$A = \{c_1, f(c_1), f^{\circ 2}(c_1), f^{\circ 3}(c_1)\}, \quad B = A \cup \{v\},$$

we get a deformation space $\mathcal{D}_f \subset \mathrm{Teich}(B)$. Its image in the moduli space is the set of conjugacy classes of quadratic rational maps with the properties described above.

Explicitly describe a simple closed curve $\gamma \subset (\mathbb{S}^2, B)$ such that a power of the Dehn twist around this curve preserves the deformation space, and its action on the set of connected components of $\mathcal{D}_f$ has an infinite orbit.

Pre-registered statement.

The problem deals with understanding the parameter space of rational functions. When the space is multidimensional, it becomes hard to study and visualize, so it's natural to restrict one's attention to subvarieties determined by certain postcritical relations. Per_n curves were introduced by J. Milnor in his study of quadratic rational functions.

These curves can be naturally embedded into modular space, with the cover being a submanifold in the Teichmüller space, called deformation space. It was an open problem whether the deformation space is simply connected, when non-empty. The first author, together with J. Kahn and N. Selinger, has shown that it is non-contractible for Per'_n curves by producing a pair of commuting Dehn twists that preserve the deformation space. At the same time E. Hironaka and S. Koch have shown that deformation space for Per'_4 curve contains infinitely many components. It was a natural idea to use the techniques of Hironaka–Koch to show that there exists a Dehn twist that permutes the components.

The problem was at the back of the mind for quite some time, we actively worked on it for several days. The solution required very good understanding of both Hironaka–Koch and Firsova–Kahn–Selinger techniques, as well as insights from the study of topology of complements to singular curves and line arrangements in $\mathbb{C}^2$, which was the main contribution of the second author.

3.8 Problem 8: Simion Filip (dynamical systems)

Let M be a compact connected smooth manifold. Let X be a smooth vector field on M.

Assume that X has the following property: If f is a smooth function, and u is a distribution, and we have $\mathcal{L}_X u = f$, then u is a smooth function. Let g_t denote the flow generated by X.

Show that the flow g_t has zero topological entropy.

Pre-registered statement.

Context. The question is closely related to a conjecture of Katok and Greenfield–Wallach (KGW), formulated independently, with different motivations, and at different points in time. The assumption in the KGW conjecture has many equivalent formulations, and agrees with the assumption in the problem statement. Vector fields satisfying this assumption are called "cohomology-free", or "globally hypoelliptic". The KGW conjecture then is that a manifold with such a vector field must be diffeomorphic to a torus, with the vector field having constant coefficients and satisfying an additional Diophantine condition.

The stated problem is the much weaker conclusion that such a flow must have zero topological entropy. This statement appeared a few times in the literature without proof, or with an incorrect proof, but is open as of now to the best of my knowledge.

In the broad classification of dynamical systems as exhibiting elliptic, parabolic, or hyperbolic behavior, the KGW conjecture aims to characterize the elliptic Diophantine ones. The problem at hand shows that such systems are not hyperbolic.

Novelty. There has been a lot of work on the KGW conjecture, with different techniques. It is known in dimension up to 3, and under various topological assumptions on the underlying manifold. Since the claim about zero entropy has appeared in the literature but no proof has been given, I assume other researchers have already tried existing techniques.

Strategy. The strategy of my proof is to apply the solvability of the cohomological equation, but to distributions instead of functions. This yields families of distributions, which control ergodic integrals but also admit useful geometric estimates. Using this, one can derive some geometric consequences, including the zero entropy property. The arguments apply equally well to flows and discrete-time dynamical systems (which is not the case for some of the other techniques in the literature).

Difficulty. I believe that once the correct direction and intermediate statements are identified, the proofs are not very difficult, for example an expert could give the proofs in a few hours. But I think that coming up with the right direction and useful intermediate statements requires some novel ideas.

3.9 Problem 9: Benjamin Dozier, Sidhanth Raman, and Carlos Serván (Teichmüller theory)

Let $\mathcal{M}_g$ be the moduli space of genus g Riemann surfaces. We consider the Deligne–Mumford compactification $\overline{\mathcal{M}}_g = \mathcal{M}_g \sqcup \partial\mathcal{M}_g$. The boundary $\partial\mathcal{M}_g$ admits a partition into strata Δ_G parametrizing stable curves with dual graph G.

Let $V_0 \subset \mathcal{M}_g$ be a (complex) smooth subvariety of (complex) dimension 2 that is totally geodesic with respect to the Teichmüller metric. Let V_t be a holomorphic family of subvarieties of $\mathcal{M}_g$ deforming V_0, for t in the unit disc $\Delta \subset \mathbb{C}$ [3]. We impose several technical assumptions on this family, namely (all closures below, e.g. $\overline{V}_t$, are taken in $\overline{\mathcal{M}}_g$):

(i) for every boundary stratum Δ_G, we have that $\overline{V}_t \cap \Delta_G$ is a (smoothly trivial) family, and

(ii) the families so produced are compatible with V_0 in the sense that if the boundary of V_0 in a smaller stratum arises from its boundary in a larger stratum, then the same holds for V_t. That is, if $\Delta_G, \Delta_{G'}$ are strata with $\Delta_{G'}$ on the boundary of Δ_G, and U is an open set in $\overline{\mathcal{M}}_g$ with

$$\overline{\overline{V}_0 \cap \Delta_G} \cap \Delta_{G'} \cap U = \overline{V}_0 \cap \Delta_{G'} \cap U,$$

then, for all t,

$$\overline{\overline{V}_t \cap \Delta_G} \cap \Delta_{G'} \cap U = \overline{V}_t \cap \Delta_{G'} \cap U.$$

[3] i.e. a complex submanifold $\mathcal{V} \subset \mathcal{M}_g \times \Delta$, with $V_t = \psi(\phi^{-1}(t))$, where $\phi : \mathcal{V} \to \Delta, \psi : \mathcal{V} \to \mathcal{M}_g$ are the natural projections; we will further assume that this gives a smoothly trivial family of submanifolds.

Note that we do *not* assume that V_t is totally geodesic for $t \neq 0$.

Now fix Γ a dual graph; we will assume for ease of notation that Δ_Γ is a product of lower complexity moduli spaces $\mathcal{M}_{g_1,n_1} \times \cdots \times \mathcal{M}_{g_k,n_k}$ [4]. For each i, let $\pi_i : \Delta_\Gamma \to \mathcal{M}_{g_i,n_i}$ be the natural projection map.

Denote by ∂V_t the closure of V_t in $\overline{\mathcal{M}}_g$ intersected with Δ_Γ. Suppose that $\dim \partial V_t = 1$ and that each ∂V_t is constant on some factor $\mathcal{M}_{g_i,n_i}$ of Δ_Γ, i.e. for some i, $\pi_i(\partial V_t) = \{Y_t\}$ for all t. Prove that the family ∂V_t is non-varying on that factor, i.e. $\pi_i(\partial V_t) = \{Y_0\}$ for all t.

Pre-registered statement.

This problem arose from a lemma in an unfinished project on rigidity of (Teichmüller) totally geodesic subvarieties of the moduli space of Riemann surfaces, joint with Carlos Serván and Sidhanth Raman. It rules out certain boundary behavior for a family of subvarieties deforming a totally geodesic subvariety. Our proof of this lemma took us several days of reasonably hard work to find and incorporated some intuition/ideas developed over several months. A recent theorem of Benirschke–Dozier–Rached, and independently Arana-Wright, gives that the boundary in any stratum is itself totally geodesic, which is the primary tool used for such things. In the situation here the boundary is just a single point, which can be considered a "0-dimensional totally geodesic subvariety," but is clearly not rigid itself. So one can't directly use McMullen's rigidity result for Teichmüller curves. We had the idea to do some other degenerations/undegenerations, and apply McMullen's result at a new boundary point. Getting this to succeed requires some additional work, and we ended up using other results of Arana-Wright that establish existence of certain types of undegenerations.

The exact hypotheses that one needs on the permissible families in order to deduce the result are still a bit unclear, but I've imposed fairly extensive assumptions in the statement to avoid this being an obstacle in the proof. The condition that the boundary stratum Δ_Γ is a product, rather than a product quotiented by a finite group, is unnecessary, but makes the result easier to state.

I'd be interested to see if there's an alternative approach.

3.10 Problem 10: Charles Doran, James M. Jones, Eric Pichon-Pharabod, and Alan Thompson (algebraic geometry)

Let H denote the hyperbolic plane lattice and let E_8 denote the negative definite E_8 root lattice.

Let $\pi\colon \mathcal{X} \to \Delta$ be a Tyurin degeneration of K3 surfaces, with central fiber $X_0 = V_1 \cup V_2$, which is H-polarized in the following sense (this corrects [6, Definition 3.20] as it appears in the arXiv version, which omits the final condition in (2)):

1. (monodromy invariance) over the punctured disc Δ^* there is a trivial local subsystem $\mathcal{L}$ of $R^2\pi_*\mathbb{Z}$ with fiber H, so that for each $p \in \Delta^*$ the fiber $\mathcal{L}_p \subset H^2(X_p;\mathbb{Z})$ defines an H-quasipolarization of X_p,

[4] for some boundary strata, we would need to quotient this by a finite group; we exclude those here

2. (positivity) there exists an effective, flat, and nef Cartier divisor A on $\mathcal{X}$ such that the restriction A_p of A to any fiber X_p over $p \in \Delta^*$ is nef and big and satisfies $[A_p] \in \mathcal{L}_p$. Moreover, the restriction of A to V_i satisfies $(A|_{V_i})^2 > 0$ for each $i \in \{1, 2\}$.

Assume that this degeneration is very general, in the sense that the image of $0 \in \Delta$ under the map from Δ to the compactified moduli space of H-polarized K3 surfaces defined by Alexeev–Brunyate–Engel [3] is not contained in any Noether-Lefschetz locus.

Let Y be a very general $H \oplus E_8 \oplus E_8$-quasipolarized K3 surface. Then there exists an elliptic fibration $\pi\colon Y \to \mathbb{P}^1$ along with an allowable loop $\gamma \subset \mathbb{P}^1$ (in the sense of [6, Definition 4.11]), such that the $H \oplus E_8 \oplus E_8$-quasipolarization is compatible with π and γ (in the sense of [6, Definition 4.21]) and such that $\mathcal{X} \to \Delta$ and $\pi\colon Y \to \mathbb{P}^1$ are a mirror pair in the sense of [6, Definition 5.1].

Pre-registered statement.

Before getting into specifics, let me just give some context on why we picked this problem. We chose a special case of a more general result we have for a much larger class of K3 surfaces, and the whole project is something none of us have ever discussed with an AI system. The result is a form of mirror symmetry which we know how to prove by very different constructions in completely different categories on two distinct spaces.

The general proof we have is constructive and exact, but involves some quite new computational software to certify. We chose the special case because it was easier and there were more obvious connections to work by ourselves and others in the literature to draw upon. In particular, the statement itself requires the AI to read Alan's recent long paper on the arXiv and digest notation and constructions that appear nowhere else. I also suspect that a sufficiently powerful frontier model might try to attack the problem by drawing on existing proofs of mirror symmetry in other settings.

In short, for a frontier model at least, we expect to learn a lot about (1) how the AI deals with a new theory tied to essentially only one complex primary reference, (2) whether it decides to use its coding abilities to do serious certified computations in its sandbox, and (3) perhaps even whether it tries to attack the problem using methods from completely different approaches to mirror symmetry.

4 AI Systems Tested

The five systems tested in the benchmark are described below.

4.1 System A: OpenAI GPT-6 Astra

System A is OpenAI's GPT-6 Astra, called once per problem through the API with reasoning effort max (the highest setting the API offers),[5] standard mode, and web search enabled.

[5]ChatGPT also offers an "Ultra" setting. Although ChatGPT lists it beside its reasoning levels, it is not a deeper level of reasoning than max: it runs maximum reasoning while delegating parts of the task to parallel subagents. It is a ChatGPT (and Codex) harness, not available through the API, and was not tested. See OpenAI's documentation on subagents and models.

Prompt: Consider the following research-level math question.

[Question]

Carefully consider the question and provide a complete and rigorous proof. Provide either complete proofs or careful citations for all intermediate statements in the proof. Work very hard to prove the question and do not return until a complete proof is achieved.

Make sure to structure the proof with careful lemmas and theorem statements. The output must be a compilable LaTeX document conforming to the standards of rigor and scholarship prevailing in mathematical literature.

4.2 System B: Anthropic Siege (workflow on Claude Opus 5.5)

System B is a Claude Code proof workflow called *Siege* which is built on Claude Opus 5.5. It is a skill from Anthropic's public math-proof plugin for Claude Code which coordinates multi-agent work over at most fourteen rounds. The run ends with a self-contained proof.md whose Status section states what is and is not proved. All sessions used Claude Opus 5.5 at high effort, with web search and fetch disabled. A runner around the skill submitted a solution only if the skill concluded with the verdict "PROVED". Accepted proofs were then typeset into LaTeX; otherwise no solution was submitted.

The Siege skill is public: `https://github.com/anthropics/claude-plugins-official/tree/main/plugins/math-proof`. Anthropic's runner around it is published in our repository at `batch-3-submissions/siege`.

4.3 System C: IMProofBench ProofCouncil

System C, called ProofCouncil, is a multi-agent harness, which calls the following models: GPT-6 Astra Pro, GPT-6 Astra, Claude Fable 5.1, Claude Sonnet 5, and Gemini 3.1 Pro (preview), with web search enabled. It was developed by a team consisting of Tim Gehrunger (lead, ETH Zurich), as well as Johannes Schmitt (ETH Zurich), David Holmes (Leiden University), Jasper Dekoninck (ETH Zurich), Yonggang Jiang (ETH Zurich), Gergely Bérczi (Aarhus University), Uri Kreitner (ETH Zurich), and Liam Price (independent researcher).

The source code for System C is available at `batch-3-submissions/proofcouncil`

4.4 System D: MerLean

System D, called MerLean, is an open-source harness for mathematical research, which calls GPT-6 Astra and GPT-6.1 Sol, with web search enabled. It was developed by Yuanjie Ren (lead, MIT), as well as Yidi Qi (Harvard), Jinzheng Li (Northeastern University), Zeru Zhu (Stony Brook University), and Kai Chau (Matthew) Kung (Boston University).

The source code for System D is available at `batch-3-submissions/merlean`

4.5 System E: Humanize (Humanfia)

System E, called Humanize, is an open-source harness for mathematical research, which calls Kimi K3 (open-weight), self-served, without web search. It was developed by Zhengyang Zhang (lead), as well as Jing Xiong, Jui-Hui Chung, Hongzhou Lin, Xingyang Li, Junxian Guo, Enrico Shippole, Mark Kim, Ligeng Zhu

The source code for System E is available at `batch-3-submissions/humanize`

5 Results and Costs

The referees' scores for the AI-generated solutions are summarized in three tables, one for each criterion: correctness (Table 2), exposition (Table 3) and attribution (Table 4). Table 5 shows the costs as well as the token usage for each system, broken down by problem.

Recall from Section 2.3 that scores on each criterion range from -1 (worst) to 1 (best). We decided to use negative numbers in our scale because spending time reading solutions that are incorrect, poorly written, or poorly attributed, can be a substantial waste of time for a human. In particular, on the correctness axis, we wanted to incentivize systems to output no answer or admit they couldn't solve the problem, rather than to give a deceptive wrong solution.

Each entry in each table is the average of the referees' scores for that submission, and the scores for Systems A through D link to a summary of referee feedback for that problem. Meanwhile, System E, which is built on an open-weight model, is discussed separately in Section 6.

Note that System B did not submit solutions to Problems 3, 4, 8, 9 and 10 (—), for which it received a score of zero.

The full referee reports for all submissions are available on GitHub.

	GPT-6 Astra		Siege		ProofCouncil		MerLean		Humanize	
Problem	$	in/out	$	in/out	$	in/out	$	in/out	$	in/out
P1	$4.64	0.15/0.06	$56	16.4/1.44	$105	7.77/0.85	$160	74.8/0.55	$288	85.8/4.12
P2	$3.75	0.10/0.06	$24	7.32/0.58	$119	8.81/1.26	$190	88.5/0.65	$375	134/1.14
P3	$3.54	0.13/0.05	$85	23.5/2.23	$171	11.8/1.50	$951	444/3.27	$740	263/2.61
P4	$3.77	0.13/0.05	$15	5.10/0.35	$105	6.30/0.98	$176	81.9/0.60	$87	27.1/1.02
P5	$3.21	0.14/0.04	$79	20.0/1.96	$67	4.79/0.52	$197	92.1/0.68	$172	59.4/0.91
P6	$3.06	0.14/0.03	$53	17.1/1.24	$107	10.8/0.73	$1,372	640/4.71	$32	9.56/0.43
P7	$3.99	0.14/0.05	$33	12.2/0.75	$145	7.77/1.52	$126	58.9/0.43	$197	56.3/3.27
P8	$2.28	0.13/0.02	$442	172/9.00	$332	36.5/2.75	$689	321/2.37	$9.60	1.95/0.31
P9	$4.29	0.14/0.06	$534	222/10.2	$594	28.9/6.51	$803	375/2.76	$196	69.0/0.79
P10	$3.81	0.12/0.05	$79	19.4/2.00	$145	13.6/1.34	$659	307/2.26	$31	9.47/0.37
Total	$36	1.32/0.47	$1,400	515/29.7	$1,888	137/18.0	$5,323	2483/18.3	$2,127	716/15.0

Table 5: Cost per problem in dollars, with input/output tokens in millions. Basis differs by system.

Notes on costs. In Table 5, costs are list-price figures drawn from each system's own records: provider bills for GPT-6 Astra and MerLean (the latter split across problems by MerLean's token log,

	System A GPT-6 Astra	System B Siege (Claude Opus 5.5)	System C ProofCouncil	System D MerLean	System E Humanize (open-weight)
P1	1.00	1.00	1.00	1.00	−1.00
P2	0.33	0.50	1.00	0.50	−1.00
P3	0.50	0.00	1.00	0.00	−0.75
P4	1.00	0.00	1.00	0.00	−1.00
P5	1.00	1.00	1.00	1.00	−0.50
P6	1.00	0.00	1.00	0.00	−1.00
P7	1.00	0.00	0.67	0.50	−0.50
P8	0.00	0.00	1.00	0.00	−0.83
P9	0.67	0.00	0.00	0.00	−1.00
P10	0.00	0.00	1.00	−0.67	−1.00
Average	0.65	0.25	0.87	0.23	−0.86
Total Correct	7	3	9	4	0

Table 2: **Correctness** scores for each system on each problem. In the row labeled "Total Correct," we indicated the number of problems on which the system score a 0.5 or better.

so per-problem figures are estimates), Siege's run log scaled to its $1,400 Anthropic bill (including the five problems where it declined to submit), and ProofCouncil's per-call cost log. Only GPT-6 Astra, ProofCouncil and MerLean used web search, and ProofCouncil's token totals understate usage because its Anthropic counts omit cache traffic. Humanize self-hosted Kimi K3 (about 400 GPU-hours, roughly $1,960), so its figures are the equivalent cost at DeepInfra's API prices ($2.85 input, $0.285 cached input and $14.25 output per million tokens, checked 2026-10-10), with a 6.8% prefix-cache rate applied uniformly across problems.

	System A	System B	System C	System D	System E
	GPT-6 Astra	Siege (Claude Opus 5.5)	ProofCouncil	MerLean	Humanize (open-weight)
P1	0.67	−0.50	1.00	0.50	−1.00
P2	0.00	0.00	1.00	0.50	−1.00
P3	0.00	—	1.00	0.00	−1.00
P4	0.67	—	0.00	1.00	0.50
P5	−0.33	−0.33	−0.33	−0.50	−1.00
P6	1.00	−1.00	0.33	−1.00	−0.50
P7	0.00	−1.00	−0.67	−0.33	−1.00
P8	0.33	—	0.67	−1.00	0.00
P9	0.00	—	−1.00	0.00	−0.50
P10	1.00	—	1.00	0.00	−1.00
Average	0.33	−0.14	0.30	−0.08	−0.65

Table 3: **Exposition** scores for each system on each problem. Empty submissions count as 0 in the average.

5.1 Problem 1

The problem concerns cohomology for the birhizomic topology on monoids. Four submissions give correct proofs along broadly similar lines to the human solution, with some useful simplifications. The differences between these submissions are largely conceptual and expository. ProofCouncil was the reviewers' unanimous favorite.

Submission A: GPT-6 Astra. All three reviewers judge the solution mathematically correct. Its strategy is close to the human proof, but a more concrete simplicial-cohomology calculation and the avoidance of an approximation of birhizomic families simplify parts of the argument. The reviewers differed about its readability: two authors familiar with the human proof found it clear, whereas another found that individual lines were verifiable without the purpose of the argument being understandable. The reviewers also differed in their rankings relative to MerLean. The submission would benefit from overviews of its longer proofs, clearer finiteness hypotheses, and more precise Stacks Project citations. These were treated as presentational concerns; the referees found no substantial attribution problem.

Average scores (C, E, A): $(1.00, 0.67, 1.00)$. Referee reports (PDF).

	System A	System B	System C	System D	System E
	GPT-6 Astra	Siege (Claude Opus 5.5)	ProofCouncil	MerLean	Humanize (open-weight)
P1	1.00	1.00	1.00	1.00	−1.00
P2	0.00	0.00	0.00	0.00	0.00
P3	0.00	—	1.00	−0.50	−1.00
P4	1.00	—	1.00	1.00	−0.50
P5	−0.67	0.00	0.67	−0.50	0.00
P6	0.00	−1.00	0.33	−0.50	−1.00
P7	0.00	0.00	0.67	−0.33	−1.00
P8	0.00	—	1.00	−0.67	−1.00
P9	1.00	—	0.00	0.00	−1.00
P10	1.00	—	1.00	0.00	−1.00
Average	0.33	0.00	0.67	−0.05	−0.75

Table 4: **Attribution** scores for each system on each problem. Empty submissions count as 0 in the average.

Submission B: Siege. Both completed reports assess the argument as mathematically correct, following essentially the same strategy as the human solution. The principal weakness is the exposition. The submission works directly in considerable generality rather than making the proof modular through reductions, and key passages require the reader to recall many symbols defined several pages earlier. One reviewer counted roughly twenty notational entities in a short proof, illustrating the burden of keeping many symbols in mind. Both reviewers qualified their confidence by noting the difficulty of checking the text, although the crucial arguments they examined were correct. The references were acceptable, and one reviewer considered them more numerous than necessary. Substantial rewriting would improve accessibility without changing the mathematical argument.

Average scores (C, E, A): $(1.00, -0.50, 1.00)$. Referee reports (PDF).

Submission C: ProofCouncil. All three reviewers found the proof correct and preferred it to the other submissions; one considered it better than the human solution. Its principal simplification is to introduce an auxiliary sign topology and reduce to sign covers, rather than the more general finite covers used elsewhere. A Mayer–Vietoris argument then replaces a more elaborate simplicial-complex calculation. These choices also improve the conceptual explanation: the submission gives a clear overview of the strategy and explains how to think about the problem. Some terminology is nonstandard, and a few inductive or technical passages require more careful wording, but the

reviewers regarded these as stylistic issues. The attributions were judged thorough and appropriate.

Average scores (C, E, A): $(1.00, 1.00, 1.00)$. Referee reports (PDF).

Submission D: MerLean. Both completed reports judge the proof mathematically correct. It follows the general strategy of the human solution but includes a simplification in Proposition 8.2 that avoids the human proof's finite-type approximation at that step. One reviewer ranked it second among the submissions. The main reservations concern exposition: the domain of a variable and the indexing of a cover are left unspecified, and the argument sometimes switches between the category of monoids and its opposite without sufficient warning. The overall strategy is readable, while individual statements can be difficult to parse. The reviewers found the attributions acceptable overall, although more precise citations and a shorter treatment of some standard material would improve the presentation.

Average scores (C, E, A): $(1.00, 0.50, 1.00)$. Referee reports (PDF).

5.2 Problem 2

The problem asks for the minimax rate of conductivity estimation in the Darcy inverse problem. Four submissions identify the expected rate, with varying degrees of rigor. ProofCouncil obtains matching bounds without logarithmic factors, improving the human upper bound. Humanize claims an incorrect, faster rate. Missing attributions recur across the otherwise promising submissions.

Submission A: GPT-6 Astra. The submission gives the expected rate up to logarithmic factors and a plausible strategy based on a plug-in kernel estimator and the Darcy inversion formula, without a preliminary estimation step. The referees disagree substantially about the completeness of its proofs. Reviewer A finds the technical arguments seemingly flawless, while Reviewer B assigns correctness 0.5 despite missing proof details; Reviewer C finds substantial gaps that prevent reliable verification, notably in the concentration estimate and the analysis near the critical point. The exposition hides the estimator and error decomposition inside technical arguments and provides too little motivation. All three reviewers identify inadequate attribution. The supplied reports leave unresolved whether the central gaps can be repaired routinely.

Average scores (C, E, A): $(0.33, 0.00, 0.00)$. Referee reports (PDF).

Submission B: Siege. All three reviewers find the estimation and proof strategy sound, with minor gaps in the justification. The upper bound uses the Darcy inversion formula and a plug-in kernel estimator without sample splitting, treating the L^2 risk directly through a different error decomposition from the human solution. The lower bound applies Assouad's lemma to conductivity perturbations, and the bounds match up to logarithmic factors. The heuristic explanation of the rate is useful, but the proofs are dense, connected results are split between lemmas, and some notation is used before it is defined. Multivariate derivatives also require correction to proper multi-index notation. The submission cites some tools correctly but omits related work and reproves established results without adequate credit.

Average scores (C, E, A): $(0.50, 0.00, 0.00)$. Referee reports (PDF).

Submission C: ProofCouncil. All three reviewers judge the mathematics essentially correct. The submission obtains the exact minimax rate without the logarithmic loss in the human upper bound, using a plug-in kernel estimator, the inverse transport formula, and a refined decomposition of the L^2 risk. It also avoids sample splitting. One reviewer singled it out as the cleanest and most impressive solution. The reviewers praise its consistent notation, explanations of the intrinsic geometry, and organization of the technical arguments; remaining presentational concerns are minor. The principal weakness is attribution: the related-work discussion does not replace citations at the point of use, including credit for Richter's inversion formula and analytical techniques related to Nickl–van de Geer–Wang. The literature discussion also omits relevant work on the Darcy inverse problem.

Average scores (C, E, A): $(1.00, 1.00, 0.00)$. Referee reports (PDF).

Submission D: MerLean. Both completed reports find the overall strategy sound and the minimax rate correct up to logarithmic factors, with minor details requiring completion. The upper bound combines the Darcy inversion formula with a plug-in kernel estimator, sample splitting, and finite-dimensional approximation. Its direct treatment of the L^2 risk and error decomposition differ from the human solution, but the extra estimation machinery makes the argument unnecessarily convoluted. The lower bound uses conductivity perturbations and Assouad's cube technique without acknowledging the latter. Views of the exposition were mixed: the related-work discussion is welcome, but some terminology is unexplained or invented. Established analytical and statistical tools, including the inversion formula, need attribution where they are used.

Average scores (C, E, A): $(0.50, 0.50, 0.00)$. Referee reports (PDF).

5.3 Problem 3

The problem asks for explicit isomorphisms relating a reduced stated skein algebra to a generalized quantum double and a quotient to the quantum group. ProofCouncil gives a complete proof; GPT-6 Astra follows a similar strategy with gaps in justification. MerLean acknowledges a partial result, while Humanize's attempted proof contains serious errors.

Submission A: GPT-6 Astra. Both completed reports find the overall strategy essentially correct and close to the human solution, but identify important steps that need more precise justification. These include the boundary-exchange relations and the use of saturated systems of arcs to obtain the required factorization. The reviewers believe an expert could repair the argument without changing its structure, and both assign correctness 0.5, although one calls for a major rewrite before journal publication. The mathematics resembles ProofCouncil's but is presented less precisely. Definitions of generators and relations are incomplete, invented theorem names obscure references, and diagrams would greatly aid verification. The bibliography also omits the relevant $n = 2$ result and needs more precise references to the skein-theoretic inputs.

Average scores (C, E, A): $(0.50, 0.00, 0.00)$. Referee reports (PDF).

Submission B: Siege. *No submission.*

Submission C: ProofCouncil. All three reviewers judge the proof mathematically correct, well written, and appropriately referenced. It follows the human strategy: present the quantum double, identify the skein algebra through its two Borel pieces, construct the algebra map, and establish bijectivity using a saturated system of arcs before passing to the quotient. One reviewer welcomes details that were omitted from the human solution. The reviewers regarded this as the strongest submission. Reviewer C suggests that its isomorphism differs from the human construction in possibly failing to preserve the Hopf structure, which the question does not require. The main expository concern is the lack of explanatory diagrams for the skein relations and mixed products. Attribution is strong overall, with only a minor omission concerning earlier work on the SL_3 case.

Average scores (C, E, A): $(1.00, 1.00, 1.00)$. Referee reports (PDF).

Submission D: MerLean. The submission explicitly acknowledges a partial result. It establishes the quantum-group presentation that forms the first and easiest stage of the common proof strategy, but does not carry out the essential skein-theoretic identification. Both completed reports find the algebraic material it establishes correct, but this is an elementary preliminary step rather than a solution of the problem. The exposition is more readable than some competing submissions, but it takes many sections to present the partial result and often leaves the reader unsure what has been achieved. Attribution is inadequate: the bibliography omits the principal source defining the relevant stated skein algebras and the existing solution for $n = 2$. The reviewers disagree about the severity of these attribution failures.

Average scores (C, E, A): $(0.00, 0.00, -0.50)$. Referee reports (PDF).

5.4 Problem 4

Submission A: GPT-6 Astra. The proof is correct, and the reviewers find it more elegant than the human solution. The exposition is acceptable, but some definitions are missing, some notation is undefined, and the presentation does not make clear which lemmas come from the literature and which are original. Attribution is satisfactory, with all the appropriate sources cited.

Average scores (C, E, A): $(1.00, 0.66, 1.00)$. Referee reports (PDF).

Submission B: Siege. *No submission.*

Submission C: ProofCouncil. The proof is correct. The solution technique is computationally intricate, leaving considerable room for error. The exposition is good, but could be substantially improved to reduce the effort required to check the argument. Attribution is satisfactory.

Average scores (C, E, A): $(1.00, 0.00, 1.00)$. Referee reports (PDF).

Submission D: MerLean. The submission proves only a trivial partial result and does not carry out the bulk of the argument. It arrives at the same definitions and proof scheme as the human solution, but fails to complete the induction. The exposition is acceptable. Attribution is good, with all the appropriate sources cited.

Average scores (C, E, A): $(0.00, 1.00, 1.00)$. Referee reports (PDF).

5.5 Problem 5

This problem turns out to admit a solution using existing work in the literature. Essentially, a solution may be found by combining a paper of Connes & Consani and a paper of Hall; the former of which is directly about hyperfields. Once these two papers have been found, no great effort is required to solve the problem, more on the level of a challenging homework problem (perhaps 2 pages once definitions have been given) than a research question. The review should be read in that light. All the correct AI solutions used the same method.

Submission A: GPT-6 Astra. All three referees find the counterexample mathematically correct. It uses the same infinite cyclic projective plane/hyperfield construction as the other successful AI submissions, which differs from the human solution. The exposition is the main complaint after attribution. There is no overview of the proof, and the infinite difference set is constructed before the reader learns why it is needed, so much of the argument is unmotivated. The hyperfield axioms are given in paragraph form, which makes them hard to check, and the notation is sometimes poor. Attributions are weak according to all referees, essentially equivalent to plagiarism. Hall's construction of an infinite difference set from a Sidon-type set is used with no credit at all. Connes–Consani is mentioned only once, cryptically, in the introduction, even though the hyperfield section follows their proof closely. Prenowitz and Singer are not acknowledged. One referee recommends a major revision that replaces the reproduced proofs with citations.

Average scores (C, E, A): $(1.00, -0.33, -0.67)$. Referee reports (PDF).

Submission B: Siege. All three referees find the solution mathematically correct. It gives the now-familiar counterexample: the Connes–Consani hyperfield of an infinite cyclic projective plane, with the plane built from a perfect difference set in $\mathbb{Z}$ by Hall's greedy method, seeded so that $0, 1, 4$ are collinear but 0 is off the line through $3, 12, 5$. Two referees find the exposition acceptable, while the third singles out the case organization as perhaps the clearest of the submissions. One referee says that the hyperfield definition is stated loosely under "Conventions," variables are not quantified, t is never defined, the induction is unclear, and the overall organization is confusing. The attribution problem is flagged by all three referees. Connes–Consani, Hall and (implicitly) Singer are mentioned only in the overview and never where their constructions and proofs are reproduced almost verbatim. That makes it hard to tell what is new, and the reproduced proofs should be replaced by citations.

Average scores (C, E, A): $(1.00, -0.33, 0.00)$. Referee reports (PDF).

Submission C: ProofCouncil. All three referees agree that the solution is mathematically correct. It answers the question in the negative by building a hyperfield from an infinite cyclic projective plane: the Prenowitz/Connes–Consani construction of a hyperfield from a projective plane, combined with Hall's 1947 extension of a partial difference set to an infinite perfect difference set in $\mathbb{Z}$. Parameters are chosen so that the cubic identity fails. One referee praises the clear strategy outline and calls the exposition and attribution the best of the submissions reviewed. All three referees, however, find the write-up too verbose. It reproduces the arguments of Hall and Connes–Consani at length instead of citing them, and the quotient-hyperfield discussion interrupts the main argument. One referee recommends a major revision that removes the reproduced proofs. Attribution is this submission's strongest category: Section 9 discusses and compares the prior literature. Referees asked for the citations to appear next to the relevant constructions and lemmas.

Average scores (C, E, A): $(1.00, -0.33, 0.67)$. Referee reports (PDF).

Submission D: MerLean. Both referees find the mathematics correct. Like the other successful submissions, it combines the Connes–Consani hyperfield from a projective plane with Hall's construction of an infinite cyclic plane. The exposition is readable but labored. Referees point to awkward wording, definitions made inline in paragraphs rather than in environments, definitions given in the wrong order, and unexplained phrases such as "cubic indices" and "one particular line." Attribution is the main weakness. The extension argument (Lemmas 5.1–5.2) is essentially Hall's Theorem 3.1. Hall is named but his paper is never cited: the submission cites Thas instead. The submission cites Connes–Consani without making clear that its construction and proofs are almost identical to theirs, and it omits Prenowitz. All references are arXiv versions rather than published versions, one of them superseded.

Average scores (C, E, A): $(1.00, -0.50, -0.50)$. Referee reports (PDF).

5.6 Problem 6

Submission A: GPT-6 Astra. The referees find the proof mathematically correct. The exposition is satisfactory overall. Attributions are mostly appropriate, but the referees identify repeated citations to secondary or expository sources instead of the original work. They also note missing attribution for the Poincaré conjecture and missing relevant literature for Lemma 6.

Average scores (C, E, A): $(1.00, 1.00, 0.00)$. Referee reports (PDF).

Submission B: Siege. Correctness of the proof was not resolved with complete confidence in the allotted time. One reviewer is confident that the proof is correct assuming the validity of Section 7. One reviewer report finds the covering-space arguments sound, but bases the favorable assessment in part on the correctness of the literature reproduced in Section 7. The exposition is poor; hypotheses are left unstated, notation is used before it is defined, arguments are convoluted, and key proofs are presented as lists of bullet points. These defects substantially obstruct verification. The attribution problems are serious. The reviewers identify Section 7 as largely plagiarized from Hom–Lidman–Zufelt's *Reducible surgeries and Heegaard Floer homology*, with AI assistance

helping to identify the similarity. One reviewer also reports an uncredited lemma of Howie, an incorrect source, and misplaced citations.

Average scores (C, E, A): $(0, -1.00, -1.00)$. Referee reports (PDF).

Submission C: ProofCouncil. This is a correct and complete solution. It carries out the covering-space and rank argument outlined above and combines it with the relevant results in the literature. The early outline of the strategy is helpful, but the referees differ substantially on readability. One finds the proof reasonably easy to follow; another reports that the proof of the main theorem and the notation require extreme care to verify. Specific complaints include forward references, imprecise language, and excessive reviewing or reproving of basic facts. One referee also faults the submission for failing to make clear that its cited literature already reduces the question to hyperbolic knots. Attributions are largely appropriate and complete, but some claims lack a specific reference and some citations point to secondary papers rather than the original sources.

Average scores (C, E, A): $(1.00, 0.33, 0.33)$. Referee reports (PDF).

Submission D: MerLean. This submission gives a correct partial solution and acknowledges that limitation. It identifies the appropriate covering space and derives the relevant inequalities, reducing the remaining question to cases addressed in the literature, but stops short of a complete proof. The exposition is poor. Statements are often vague or imprecise, and referees report being able to discern the intended meaning largely because they already know the argument. The relevance of Section 8 is questionable; one reviewer criticizes an unmotivated discussion of a formal difference obstruction and an irrelevant final discussion of variations on the problem. Attributions are deficient throughout, with missing or misplaced references and standard results reproved without acknowledgment of their status or sources.

Average scores (C, E, A): $(0.00, -1.00, -0.50)$. Referee reports (PDF).

5.7 Problem 7

The problem has two parts, with the substantial part requiring an infinitude result. All of the submitted proofs involve substantial computations which made it difficult for the reviewers to determine correctness with full confidence in the allotted time. The reviews therefore distinguish between identifying a strategy that appears viable, and establishing that the calculations needed to complete the strategy have been carried out correctly.

Submission A: GPT-6 Astra. The reviewers believe that this proof follows a strategy that should work; whether every step is completely correct, the reviewers were unable to check given the allotted review time due to the heavy computational nature of this proof. The exposition was satisfactory; the reviewers agree that the exposition of GPT-6 Astra was the best among the 5. Attribution was not particularly thorough, but was overall satisfactory.

Average scores (C, E, A): $(1, 0, 0)$. Referee reports (PDF).

Submission B: Siege. The reviewers believe that this proof follows a strategy that should work and is a novel approach compared to the others. The reviewers were unable to check given the allotted review time due to the heavy computational nature of this proof and quality of exposition, whether every step is completely correct. The exposition is poor: it makes a poor choice of notation, it is overly detailed on pointless numerics and calculations, and it has a poor structuring of arguments (within proofs and between proofs). This all comes to make it hard to determine if the overall solution is correct. In short, the proof seems to identify correctly what should be done, but the reviewers cannot determine if it was done. Attribution was poor.

Average scores (C, E, A): $(0, -1, 0)$. Referee reports (PDF).

Submission C: ProofCouncil. The reviewers believe that this proof follows a strategy that should work; whether every step is completely correct, the reviewers were unable to check given the allotted review time due to the heavy computational nature of this proof. The exposition notably more convoluted than the exposition of GPT-6 Astra. Attribution was not particularly thorough, but satisfactory.

Average scores (C, E, A): $(2/3, -2/3, 2/3)$. Referee reports (PDF).

Submission D: MerLean. MerLean's method of proof introduces a novel idea compared to the other solutions. The exposition is quite poor, however, and so reviewers were unable to determine with full confidence that the solution is correct. If correct, this proof is novel and interesting. Attribution is somewhat problematic: several results that exist in the literature are derived from scratch without citing the literature. Several citations are not in the proper place.

Average scores (C, E, A): $(1/2, -1/3, -1/3)$. Referee reports (PDF).

5.8 Problem 8

Submission A: GPT-6 Astra. GPT-6 Astra's solution does not solve the problem, and is honest about the fact that it doesn't solve the problem. Its main content is a reduction between the flow and diffeomorphism versions of the question. The referees described this as a standard suspension argument already in the literature (including work by Rodriguez-Hertz–Rodriguez-Hertz), which makes little progress toward a solution. Thus, it received a correctness score of 0. Opinions on the exposition were mixed, from neutral to good; one referee remarked that it reads easily mainly because the mathematics is trivial. Attributions were judged roughly neutral, with one referee noting that the known reduction is not attributed.

Average scores (C, E, A): $(0.00, 0.33, 0.00)$. Referee reports (PDF).

Submission B: Siege. *No submission.*

Submission C: ProofCouncil. All three referees judged ProofCouncil's solution fully correct (unanimous 1), with no mathematical errors found; Referee 2 noted that the argument differs substantially from the author's own solution, and Referee 1 observed that it actually proves something stronger than the problem asks (that the distribution is not a measure), which should have been stated as such. Attributions were also rated unanimously as fully appropriate. Exposition drew the only criticism: Referee 1 found it imprecise and spending too much space on routine material, and several referees independently asked for a picture to illustrate the test-function construction in Section 8 (including how Saussol's Lemma 4.1 is used), and for a certain proposition to state the polynomial lower bound that is actually proved rather than the weaker exponential one. Referee 3 added minor suggestions on notation, more modern references for stable manifolds, and reminders about the choice of non-atomic and non-periodic points.

Average scores (C, E, A): $(1.00, 0.67, 1.00)$. Referee reports (PDF).

Submission D: MerLean. MerLean's solution openly says that it does not solve the problem, so all three referees gave correctness a 0 (for an acknowledgment of the failure). The referees were critical of the rest of the solution: it spends its effort re-deriving, from scratch and unsuccessfully, facts already in the literature (Greenfield–Wallach, Katok, Forni) on smooth invariant measures, and it includes textbook ergodic-theory bounds that have nothing to do with global hypoellipticity. They found the exposition poor: claims are deferred without saying where they are proved, others are treated as common knowledge, and it is unclear how the extra material helps toward a solution. One referee referred to the exposition in our discussions as "a big circle to nowhere." Minor wording problems were also noted (an unnamed Lyapunov function, "probability torus"). Two of the three referees considered the missing references a serious attribution problem.

Average scores (C, E, A): $(0.00, -1.00, -0.67)$. Referee reports (PDF).

5.9 Problem 9

The problem concerns rigidity of a boundary factor under deformations of a totally geodesic surface in moduli space. GPT-6 Astra finds the central pinching–unpinching idea of the human proof, though its presentation is considerably more complicated. ProofCouncil and MerLean give acknowledged partial results under additional assumptions; Humanize's claimed partial progress rests on false inputs.

Submission A: GPT-6 Astra. The reviewers agree that the submission captures the essential mathematical idea. Its pinching–unpinching argument constructs two boundary paths with a common stable limit and uses smoothness to force the relevant factor to remain constant. This is the central mechanism of the human proof, and its recovery impressed the reviewers. Nevertheless, two assign correctness 0.5 because some statements and intermediate justifications require correction or clarification. All three reviewers found the exposition difficult to follow. The submission fails to apply its key idea uniformly, instead making unnecessary reductions and reproducing McMullen's rigidity argument at length. Standard facts receive excessive attention while important steps remain

compressed. The references used are appropriate, although broader context and a substantial reorganization would make the proof far easier to understand.

Average scores (C, E, A): $(0.67, 0.00, 1.00)$. Referee reports (PDF).

Submission B: Siege. *No submission.*

Submission C: ProofCouncil. All three reviewers treat the submission as an acknowledged partial result. It handles boundary strata with at least two nodes and discusses a nonarithmetic case, but the remaining one-node arithmetic case is addressed only after adding an unproved lifting hypothesis. Despite this partial coverage, the core case remains unsolved. The submission misses the efficient pinching–unpinching argument and pursues elaborate algebraic-geometric constructions and sewing obstructions whose relevance is unclear. One reviewer sees a possible hint of the right idea, but it is buried in unnecessary material. The exposition is uniformly rated poor. Its citations are generally correct where supplied, but it omits relevant work of Benirschke–Dozier–Rached and inadequately credits some underlying ideas.

Average scores (C, E, A): $(0.00, -1.00, 0.00)$. Referee reports (PDF).

Submission D: MerLean. All three reviewers regard the submission as an acknowledged partial result. It proves the desired conclusion under four sets of additional hypotheses, but does not show that those hypotheses follow from the problem's assumptions. In particular, it does not use total geodesicity of the central fiber to bridge the central gap. Despite this limitation, it was judged the best written of the four submissions. The reviewers find no obvious major errors in the conditional arguments, and one reduction involving several nodes is correctly executed, but they see little progress on the difficult case of the original question. The exposition is reasonably clear about what remains unresolved, though verification still requires effort. Relevant work of Benirschke–Dozier–Rached is missing, and some claims lack adequate attribution.

Average scores (C, E, A): $(0.00, 0.00, 0.00)$. Referee reports (PDF).

5.10 Problem 10[6]

Submission A: GPT-6 Astra. All three referees report that GPT-6 Astra honestly says explicitly that it cannot give a complete proof. It checks a few easy things, which one referee estimated was about 10% of a solution. Another referee called its progress "essentially empty." Under the rubric, this earns an average correctness score of 0. The referees agree that the material it does contain is sound, clearly explained, and appropriately short. They also agree that its attributions are appropriate.

Average scores (C, E, A): $(0.00, 1.00, 1.00)$. Referee reports (PDF).

[6] In our first run of Problem 10, an inconsistency between two sources in the literature was identified. The statement was subsequently corrected on October 4, 2026. Only the corrected run is released.

Submission B: Siege. *No submission.*

Submission C: ProofCouncil. All three referees judge ProofCouncil to be an essentially flawless proof. One referee considers it better than the human-written solution, because it avoids a hard explicit computation by splitting a large unimodular lattice into smaller unimodular summands in a natural way. Another referee notes that the model correctly recognized that unimodularity of the lattices simplifies the conjecture substantially, and that, alone among the submissions, it cites primary sources for standard material. The criticisms are minor: notation that is reused inconsistently (e.g. a, b with two different meanings), a step in the construction of ψ that could be referenced more explicitly, and some citations to recent papers for results originally due to Kulikov, Friedman, or Shepherd-Barron, plus citations without section or theorem numbers (R2, R3).

Average scores (C, E, A): $(1.00, 1.00, 1.00)$. Referee reports (PDF).

Submission D: MerLean. MerLean claims a counterexample to the statement, but all three referees agree that the counterexample is not valid. The referees felt that the solution appeared deliberately deceptive: it depends on a sign convention taken from arXiv version 2 of Giovenzana–Thompson [GT] which was corrected in version 3. Two referees mention that the model knowingly used the outdated version: it cites arXiv versions very precisely and adds disclaimers that its conclusions hold only for the definitions exactly as written. The third referee points out that the model should have caught its mistake; the erroneous sign can be checked directly against the pseudolattice constructions in Giovenzana–Thompson, using an explicit integral matrix calculation. It is telling that MerLean explicitly recognizes that its result contradicts Giovenzana–Thompson's established degree formula, but attributes the discrepancy to conventions without checking the simplest examples of those conventions. The referees agree that the construction of the degeneration is essentially correct and that the computations are clearly written, but finding the source of the error took the referees a lot of time. That is why the referees' exposition assessments differ widely. On attribution, all three object to the reliance on the superseded version of [GT] and the lack of primary sources (Kulikov, Persson–Pinkham, Friedman, Shepherd-Barron, Looijenga, Gross–Hacking–Keel, Kondō).

Average scores (C, E, A): $(-0.67, 0.00, 0.00)$. Referee reports (PDF).

6 Open-weight submission: Humanize

We have received considerable interest in how open-weight models compare to the frontier models. System E, Humanize, built by the Humanfia team, is the only official open-weight submission in this batch: it is a harness built on Kimi K3. It was refereed through the same procedure as the other four systems. We summarize the referees' assessments below. What we found is that there is a substantial gap between the performance of open-weight models and the frontier models.

In order to have a more grounded comparison, we asked one reviewer for each problem to provide a comparison between the Kimi K3 harness results and the performance of GPT-5.2 Pro (which was released in December 2025). We provide a brief summary of their comments, as well as discussing additional open-weight models which were tested.

6.1 Referee assessments

Problem 1

All three reviewers found that the submission fails to solve the problem. It announces a restriction to finitely generated monoids, but then applies that restriction to the auxiliary monoids used to define infinitesimal maps, thereby trivializing the topology and invalidating its argument. Even in this restricted setting, the proof confuses faces, quotients, and different notions of generation, and contains numerous false statements. Some elementary calculations are sensible, but the submission ultimately attempts to prove the wrong statement and confuses the objects and maps involved. Its acknowledgment of limited scope does not resolve these errors. The exposition introduces undefined or misleading terminology, and there is no bibliography or substantive attribution.

Average scores (C, E, A): $(-1.00, -1.00, -1.00)$. Referee reports (PDF).

Problem 2

Reviewer C, who supplied the only completed written report, finds that the main result is incorrect. Its least-squares and multiresolution approach claims a rate with denominator $2\alpha + d + 2$ where $2\alpha + d + 3$ is required. The analysis recognizes degeneracy of the transport field near the origin but fails to account for its effect on the risk, missing the central difficulty of the problem. There are further inconsistencies in the wavelet construction and insufficient justification of key lemmas. The exposition makes the error difficult to isolate, and the submission provides no substantive account of related work. The different overall approach does not produce a valid solution.

Average scores (C, E, A): $(-1.00, -1.00, 0.00)$. Referee reports (PDF).

Problem 3

Although it follows the same broad plan as the successful proofs, the submission contains several serious algebraic errors and does not justify the essential presentation of the skein algebra. The completed reports identify incorrect root pairings, inconsistent cross-relations, and missing defining relations. In one calculation, an incorrect expression is effectively replaced by the desired formula without a valid argument. Claims of computational verification do not supply the missing proof. The exposition is exceptionally difficult to follow, with baffling equation labels, unstable notation, and repeated forward references. There is no bibliography, and the few inline citations fail to credit the relevant skein-algebra literature.

Average scores (C, E, A): $(-0.75, -1.00, -1.00)$. Referee reports (PDF).

Problem 4

The submitted proof reaches the wrong conclusion, so it is incorrect. As far as one reviewer can tell, there is a single fatal mistake; the subsequent reasoning leads to excessively strong and clearly false conclusions. The reviewers differ in their assessments of the exposition: it permits the mistake to be identified, but doing so requires significant effort. Attribution is poor: the bibliography is empty.

Average scores (C, E, A): $(-1.00, 0.50, -0.50)$. Referee reports (PDF).

Problem 5

Both referees find that the submission does not solve the problem. It proves only partial results and special cases and reduces the question to another statement that is no easier. It never finds the counterexample that shows the implication fails in general. Both referees gave the exposition the lowest score, citing strange and encumbered language. There are no citations. One referee thinks this is understandable given the arguments, but notes that the definition and construction of quotient hyperfields should still be cited.

Average scores (C, E, A): $(-0.50, -1.00, 0.00)$. Referee reports (PDF).

Problem 6

Referees find that the proof is incorrect. The entire argument depends on Theorem 3.7, which the referees identify as hallucinated and fatal to the main argument. One reviewer report notes that the cabling conjecture is not known for L-space knots. The exposition is poor and devotes excessive space to material already familiar to specialists. There is no bibliography, although author names and years occasionally appear inline, and the referee reports false or missing attributions.

Average scores (C, E, A): $(-1.00, -0.50, -1.00)$. Referee reports (PDF).

Problem 7

The problem asks to prove two things. The first part is the easier part, and Humanize's submission proves this part correctly. The second part is the more crucial one, and asks to prove that something is infinite. Humanize ends up proving that the object is at most 4, and so the proof contains at least one fatal error. The exposition is sufficiently poor that it is not clear where the major error occurred. Attribution is very poor; there are zero citations. Many calculations which are already well-known in the literature are re-derived from scratch, without reference to any literature.

Average scores (C, E, A): $(-0.50, -1.00, -1.00)$. Referee reports (PDF).

Problem 8

The referees agreed that Humanize's solution is not a valid proof: its key step, Theorem 4.1, is presented as a known result but is actually an open conjecture, much stronger than the problem, from which the problem follows trivially. One referee described the cited statement as entirely hallucinated from the reference, so attribution received the lowest possible score from all three referees. Exposition was considered readable but unremarkable, with one referee commenting that it is easy to read mainly because the mathematics is trivial once the conjecture is assumed.

Average scores (C, E, A): $(-0.83, 0.00, -1.00)$. Referee reports (PDF).

Problem 9

Both completed reports find that even the claimed partial progress is incorrect. The submission assumes an unproved input and then relies on false assertions about the existence and classification of totally geodesic subvarieties. These are central mathematical inputs, rather than peripheral errors. Numerous readily identifiable mistakes undermine the argument. Attribution is also seriously defective: a Mirzakhani–Wright reference is given with a fabricated author name and incorrect publication details, other claimed inputs are unsupported or hallucinated, and recent work directly relevant to the problem is omitted. The submission's acknowledgment that it cannot complete its method does not salvage the false preliminary claims. The reviewers differ only about whether the exposition is followable with effort or very poor.

Average scores (C, E, A): $(-1.00, -0.50, -1.00)$. Referee reports (PDF).

Problem 10

Both referees for Humanize find that it makes no real progress and is seriously flawed. Most seriously, the conclusion rests on hallucinated "Theorem B" and "Theorem C" of Giovenzana–Thompson, which do not exist; the relevant statement there is a conjecture. Both referees give the lowest correctness and attribution scores, citing fabricated theorems and misattributed definitions. One referee points out that it is very hard to read despite a readable outline. Only one of the two referees provided an exposition score.

Average scores (C, E, A): $(-1.00, -1.00, -1.00)$. Referee reports (PDF).

6.2 Comparison with an older closed model

We did not aim to have any official assessment of the performance of older models, but the substantial gap between the open-weight entry and the current frontier led us to pursue a comparison with an older model. For each problem, we asked a reviewer who had already reviewed the open-weight answer to compare it to an answer provided by GPT 5.2 Pro, an OpenAI model released in December 2025, which was run at the highest reasoning effort with web search enabled.

The reviewers were not asked for a full report, so the comments are informal. The following quotes are representative:

- "It has a lot of hallucinations and invented terminology, and it's pretty hard to tell what (if anything) is going on."
- "I would wager that GPT 5.2 could be a helpful assistant on this problem if it had some human prompting to about where it misunderstood the question. I'm less confident that [Humanize] could be used the same way, because of the amount of hallucination."
- "GPT 5.2 Pro's solution seems slightly better in terms of correctness . . . only due to the fact that it does not arrive at an obviously wrong conclusion, but the argument breaks apart under more careful scrutiny."

- "I think the two solutions are therefore comparable in terms of strategy and correctness."
- "A direct comparison between [Humanize] and this solution is not really possible, as the approaches are too different, and neither is effective or correct."
- "It could also be the case that [GPT 5.2 Pro] is more readable mostly because it doesn't do any of the tedious work required to prove the things it needs to, and [Humanize] at least attempts to do so."
- "[GPT 5.2 Pro] didn't even try to solve the problem!"

6.3 Further open-weight runs

For a broader perspective on open-weight model capabilities, we tested five open-weight models (Table 6) via API calls through OpenRouter, once with web search and once without. We used the same prompt as in our test of GPT-6 Astra, selected the highest reasoning effort offered, and set the temperature to each provider's default.

Table 6: Costs of single-call runs of open-weight models

Model	Provider	Cost, 10 problems (search / none)	Answers produced
Kimi K3	Together	$6.20 / $7.84	20 of 20
DeepSeek V4 Pro	Cloudflare	$2.11 / $2.22	20 of 20
DeepSeek V4.1 Flash	Together	$0.47 / $0.36	20 of 20
GLM 5.3	Together	$9.29 / $10.72	19 of 20
Qwen 3.8 2.4T	Together, Modal	$5.02 / $6.61	19 of 20

A quick automated review of these solutions indicates that none would receive a positive score, and none were passed on to a human reviewer.

7 Conclusion and takeaways

In this third batch of First Proof, we tested five AI systems on ten research-level problems in fields ranging from geometry and topology to statistics and dynamics; these problems were all known to have solutions of at most 12 pages. These problems came from the research of mathematicians who had solved these problems in their research but had not yet publicized the solutions in any way. The problem writers met at Harvard's Center for Mathematical Sciences and Applications on August 13–14, 2026, in order to refine their problems and do preliminary tests on them.

For the first time, we tested a system built on an open-weight model, Kimi K3, released by Moonshot AI earlier this summer. We also tested two harnesses designed by academic teams, plus one harness provided by Anthropic, as well as OpenAI's GPT-6 Astra. All testing took place between October 2 and 5, 2026, with 24 hours for each system and no human input.

On October 8 and 9, thirty expert reviewers met at Carnegie Mellon's Institute for Computer-Aided Reasoning in Mathematics (ICARM), to grade the anonymized solutions for correctness, exposition, and attribution of prior work. Since reading wrong solutions is a waste of time for humans, this batch introduced negative scoring to incentivize systems to output no answer or an acknowledged partial solution, rather than incorrect solutions that are claimed to be correct. As described in Section 2, the AI-generated solutions were assessed by expert referees not only for mathematical correctness, but also for the quality of their exposition and of their attribution of prior ideas (which includes citations to the existing literature). Results and scores are given in Section 5.

Our main takeaways from Batch 3 are as follows.

Frontier systems have improved. The ProofCouncil harness and GPT-6 Astra each answered two more problems correctly than in the previous batch, reaching 9 of 10 and 7 of 10 respectively. ProofCouncil and GPT-6 Astra cost $1,888 and $36 respectively (versus $3,200 and $117 in Batch 2). Each of the 10 questions in the benchmark was solved correctly by some system, setting aside citation and exposition issues.

Open-weight systems are far behind the frontier models. Every submission by the open-weight system contained major errors, at a total cost of about $2,000. Reviewers compared it to older frontier models and found it generally less capable than the leading frontier model of December 2025 (GPT-5.2 Pro). **We urge the community to immediately invest in improving open-source/open-weight models.**

There is a substantial cost for improved performance. By leveraging more compute, it is possible to significantly improve on the performance of a single query to GPT-6 Astra – but this comes at significant cost, leading to concerns that access to computational resources may become a barrier for basic research in mathematics.

Attribution still lags. In terms of attribution, the best AI systems have improved relative to Batch 2, but there are still instances where they fail. For instance the Astra solution to Problem 5 contained material that essentially constituted plagiarism. Only 4 of the 10 GPT-6 Astra solutions were judged to have adequate citations; the ProofCouncil harness, which uses Astra as a base model, was able to improve this to 7 out of 10. Moreover, it was clear that LLMs are much better at citing and learning from references that are posted to the arXiv, as compared to older published papers which are not on the arXiv. Some systems provided citations in a way that was not very helpful to the reader – for instance, putting no citations in the main body of the solution, only buried in a single paragraph that appeared at the very end or very beginning of the document.

Exposition was generally weak. In terms of exposition, even the top system ProofCouncil got full marks on only 4 solutions, while a third of its responses received negative average scores. Sometimes a solution was not mathematically correct, but we still awarded full marks on exposition – for instance, Astra's solution to Problem 10, or MerLean's solution to Problem 4. These were instances where the AI system provided some material that did not constitute a solution; but the provided

material was explained nicely. In terms of failures, many solutions were overly wordy, used awkward notation, or had problems with organization. Some solutions went "around in circles attempting to prove known statements by overly complicated means," for instance MerLean's solution to Problem 8.

Harnesses can lead to unnecessary failures. While harnesses have the potential to greatly improve performance of the base model, they can also make it worse, or exhibit "devious" behavior. For Problem 9, ProofCouncil got stuck in a loop and destroyed what would have been a correct solution produced by Astra. For Problem 10, MerLean "cheated" by using a broken definition from an older version of one of the arXiv references it was using.[7]

Future work

Going forward, we would like to put more emphasis on testing of open-weight models. We would also like to test more open-ended mathematical questions. Batch 4 of First Proof will take place from October 2026 to January 2027, with results released in the first half of January.

8 Supplementary documents

We are hosting supplementary documents on the First Proof GitHub page. *Throughout the repository, as in this report, the problems are numbered from* 1 *to* 10 *in reverse alphabetical order of the authors' last names.*

- The human-generated solutions to the problems
- The AI-generated solutions to the problems
- The referee reports on the AI-generated solutions
- The logs from the AI systems
- Source code for the harnesses created by the academic teams

References

[1] Mohammed Abouzaid, Andrew J. Blumberg, Martin Hairer, Joe Kileel, Tamara G. Kolda, Paul D. Nelson, Daniel Spielman, Nikhil Srivastava, Rachel Ward, Shmuel Weinberger, and Lauren Williams. First proof, February 2026. arXiv preprint `https://arxiv.org/abs/2602.05192`.

[2] Mohammed Abouzaid, Nikhil Srivastava, Rachel Ward, and Lauren Williams. First proof second batch, 2026. arXiv preprint `https://arxiv.org/abs/2606.18119`.

[7] This definition was corrected on the arXiv but MerLean chose to work with the older version of the paper.

[3] Valery Alexeev, Adrian Brunyate, and Philip Engel. Compactifications of moduli of elliptic K3 surfaces: Stable pair and toroidal. *Geometry & Topology*, 26(8):3525–3588, 2022.

[4] Manon Bischoff and Joseph Howlett. AI solves 79-year-old mathematical mystery. Scientific American, https://www.scientificamerican.com/article/ai-solves-79-year-old-math-mystery-of-six-dimensional-spheres/, September 2026.

[5] Albert Cohen, Ronald DeVore, and Christoph Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Analysis and Applications*, 9(1):11–47, 2011.

[6] Luca Giovenzana and Alan Thompson. Degenerations and fibrations of K3 surfaces: Lattice polarisations and mirror symmetry, May 2024.

[7] Konstantin Kakaes. AI has solved one of math's $1 million Millennium Prize problems. Quanta Magazine, https://www.quantamagazine.org/ai-has-solved-one-of-maths-1-million-millennium-prize-problems-20260908/, September 2026.

[8] Kimi Team, Moonshot AI. Kimi k3 technical report. Technical report, Moonshot AI, 2026.

[9] Melissa Lee. 'hello there the jacobian conjecture is false thanx': why a tiny social media post has mathematicians rethinking AI. The Conversation, https://theconversation.com/hello-there-the-jacobian-conjecture-is-false-thanx-why-a-tiny-social-media-post-has-mathematicians-rethinking-ai-283883, July 2026.

[10] Moonshot AI. Kimi k3 model weights. Hugging Face, https://huggingface.co/moonshotai/Kimi-K3/tree/f831ab66814297da540d832a5235f8e904f29d06, 2026. Revision f831ab66814297da540d832a5235f8e904f29d06. Accessed October 2026.

[11] OpenAI posts 700 maths preprints online: mathematicians are up in arms. Nature (news), https://www.nature.com/articles/d41586-026-03196-8, October 2026.

[12] Richard Nickl. *Bayesian Non-linear Statistical Inverse Problems*. Zurich Lectures in Advanced Mathematics. EMS Press, Berlin, 2023.

[13] RadixArk. Kimi-k3-dspark draft model. Hugging Face, https://huggingface.co/RadixArk/Kimi-K3-DSpark, 2026. Accessed October 2026.

[14] Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[15] William P. Thurston. On proof and progress in mathematics. *Bulletin of the American Mathematical Society (N.S.)*, 30(2):161–177, 1994.

Email addresses

[1] abouzaid@stanford.edu
[2] nikhil@math.berkeley.edu
[3] rward@math.utexas.edu
[4] williams@math.harvard.edu
[5] herr@vt.edu
[6] jonathan.wise@colorado.edu
[7] markus.reiss@hu-berlin.de
[8] sven.wang@epfl.ch
[9] asikora@buffalo.edu
[10] siebert@math.utexas.edu
[11] tuongle@princeton.edu
[12] chayiml@princeton.edu
[13] gordon@math.utexas.edu
[14] tanyaf@math.ksu.edu
[15] mmazin@math.ksu.edu
[16] sfilip@math.uchicago.edu
[17] benjamin.dozier@cornell.edu
[18] svraman@uchicago.edu
[19] cservan@umich.edu
[20] charles.doran@ualberta.ca
[21] jonesj@mathematik.uni-marburg.de
[22] eric.pichon-pharabod@maths.ox.ac.uk
[23] a.m.thompson@lboro.ac.uk