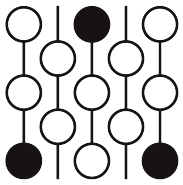
Third Batch Call for Submissions

First Proof Editorial Board

August 25, 2026



First Proof will conduct its Third Batch Benchmark during August–October, 2026. The benchmark will test commercial and non-commercial AI systems on ten naturally occurring research math problems which have been solved by human mathematicians but whose solutions have not been made public. The protocol for the benchmark will be the same as that for the second batch, with three differences, noted in red in this table:

	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
Grading		
Process	Double blind review with ∼30 expert referees	Double blind review with ∼30 expert referees
Grading 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 Benchmark.

Timeline and Deadlines. If you are interested in having your AI system tested in the benchmark, please fill in this Google Form `https://forms.gle/vqWbRsfN17vbePvy8` by **September 10, 2026**. Due to logistical and resource constraints, we will only be able to test 4–8 AI systems. We will notify you if we have decided to include your system by **September 14**. The deadline for submitting source code for your system, if accepted, will be **October 1**. The results of the benchmark will be published on **October 14, 2026**.

For a full description of the benchmark methodology, see Section 2 of https://arxiv.org/pdf/2606.18119. Below, we highlight some key points which may inform your decision to participate.

Difficulty of Problems. In addition to the increased page length, there is one significant difference in the problem selection process for Third Batch compared to Batch 2: problems which can be one-shotted during preliminary testing by GPT 5.6 Sol Pro using this prompt https://1stproof.org/assets/docs/openai-system-c-prompt.txt will not be included in the benchmark. The rationale for this is that the purpose of the benchmark is to probe the boundary of AI's capabilities in research math, and excluding problems solvable by the state of the art model in August will focus the benchmark in on problems which may be challenging for models in October. We will report the number of problems discarded in this manner from the initial pool of problems considered.

Eligibility of AI Systems. An AI system will be eligible for the benchmark if it is an *open-source harness which calls publicly available models via API*. Publicly available means that the model should be accessible to any member of the public on October 7, 2026 (one week before the results are published). A single prompt is viewed as a special case of an open-source harness.

Testing Protocol. The source code/prompt/API keys for each AI system must be provided to First Proof by October 1, 2026. The code must take as input the LaTeX source of the ten problems, run for a wall clock time of at most 24 hours, and produce ten LaTeX documents containing solutions of at most 16 pages each. First Proof will run each system on its own AWS account. Detailed technical specifications will be provided to each team whose system is accepted for testing.

Grading. As is noted in Table 1, instead of the one-dimensional "accept/major revisions/minor revisions/reject" scoring system used in Batch 2, human graders will, in addition to providing a detailed review of each AI solution, assign separate scores for correctness, exposition, and attribution. There will be negative scoring for wrong (as opposed to blank) solutions on the correctness axis.

Publication of Results. After the solutions generated by AI systems are graded by humans, the editorial board will publish each problem together with the name of its author, the author's solution, and the author's descriptive statement. The board will also publish the source code for all tested systems, the AI solutions, the referee reports, and the editorial board's scores for each solution.

Funding. Batch 3 will be supported in part by unrestricted donations from commercial companies, some of which are submitting systems for testing. Funding received from such companies is used exclusively for scientific and outreach activities, such as creating the questions and grading the AI generated solutions, and is never used to pay members of either board. We are registered as a 501(c)(3) corporation and will release full reports of our expenses.