Second Batch Announcement

First Proof Editorial Board

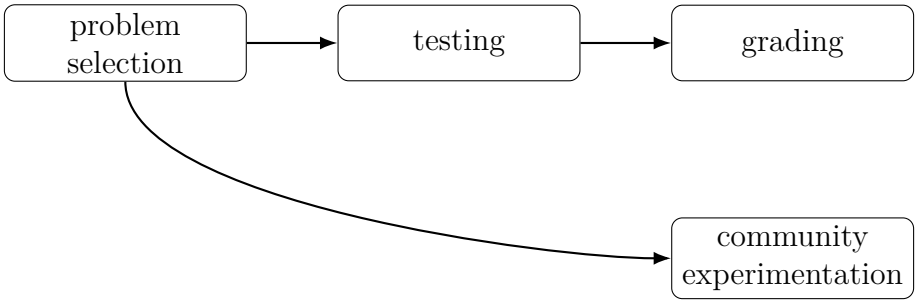
March 14, 2026

1 Introduction

The first batch of First Proof [1] was an informal collaborative experiment, consisting of 10 problems that could be used to measure the ability of AI systems to provide proofs of mathematical statements. More specifically, these problems were statements which had come up naturally in the mathematical research of the authors – but whose proofs could not be found simply by doing a literature search. 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.

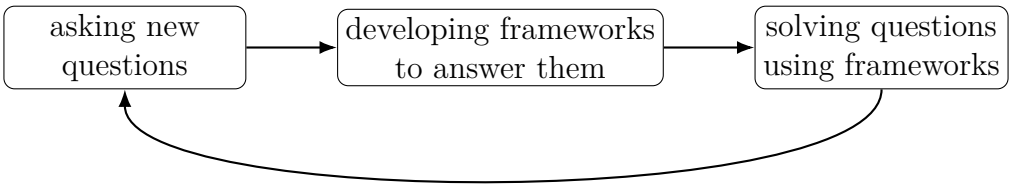
This document describes our plan for a second batch of problems, which will be created, tested, and graded from March to June 2026, and which will be overseen by the First Proof Foundation. The guiding principle of this design is transparency. This batch will be designed as a formal **benchmark**.

It will also include a separate round of informal **community experimentation**, in which a set of problems is made available to the interested public, with solutions provided after a few days, followed by an open discussion.



Before discussing more details, it is worth taking a moment to discuss what mathematics is and what mathematicians do. Mathematics is an inherently creative discipline, in which mathematicians push the boundary of mathematics by asking new questions and developing frameworks to answer these questions. Henri Poincaré described mathematical creation as a largely unconscious imaginative process, guided by aesthetic selection: "It is by logic that we prove, but by intuition that we discover." Indeed, the work of discovering novel and interesting

conjectures, as well as the right conceptual definitions and frameworks for approaching such conjectures, is at least as important as the work of proving such conjectures. One can make an analogy with art: the first step is to conceive of the idea; only having done so can one execute the work itself.



At the moment, 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 second batch, we will thus, as in the first batch, restrict our attention to assessing mathematical proofs that AI systems create. As in our initial experiment, we will collect solved but unpublished mathematical problems from human mathematicians, which we will pose as a challenge to AI systems to prove. Informed by the conversations sparked by [1], we will implement a more elaborate process, described in the remainder of this document, both for selecting the questions and assessing answers produced by AI systems.

2 Selection Process (March to May, 2026)

We will solicit questions from mathematicians representing a wide range of mathematical fields as well as geographic locations (mostly in the United States for logistical reasons). We aim to obtain mathematical questions which involve a nonstandard insight to answer, have a proof known to the mathematician of at most 8 pages which has not yet appeared on the internet or in any public forum, and will remain this way until June 2026.

We will ask that the authors:

- not discuss solutions of candidate questions with any AI system before the questions are announced, except in a controlled environment;
- disclose any employment or consulting work for AI companies in the last 3 months, and commit not to consult with new companies in the next 6 months. We will not accept submission from AI companies which are conflicted with our authors (or co-authors).
- produce a document of at most 8 pages which consists of a complete proof, with accurate references to the mathematical literature;
- provide a document, written for a general audience of mathematicians, describing the (i) novelty of the argument, (ii) the general strategy of the proof, and (iii) an estimate of the difficulty of the problem, e.g. via the time that it has taken to solve the problem.

Since we will require authors not to test questions on AI systems, we will test questions on zero data retention models to perform a literature search aiming to identify whether proofs of equivalent problems have appeared in the literature, and whether there is a standard approach to a problem.

We will have all solutions go through a first round of refereeing to check correctness. Out of the problems which pass this round, we will select **10 problems for the formal benchmark** and **a smaller and separate set of problems for the informal community experiment** 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. We will report the number of problems which are rejected through this process.

3 Benchmark: Testing (Late May to early June, 2026)

The editorial board will assess the publicly available systems with leading performance in research mathematics at the time via API (currently, Gemini 3.1 Deepthink and OpenAI GPT 5.4 Pro).

One particular design decision that we shall implement is that AI systems will only have **one shot** to answer each question. While it may be possible that, in practice, AI will be most useful to mathematicians as part of a dialogue in which the mathematicians give feedback to the model, we do not presently have an objective way to assess the contribution of an AI model to a solution which is produced as a collaboration between the AI model and a human who may provide additional ideas.

Subject to availability of funds, we will extend testing to AI systems that are produced by commercial and non-commercial entities that meet the following criteria:

- Preliminary analysis suggests that performance is at least comparable to that of the leading public models;
- We have a near guarantee that the outputs of the system are autonomously generated. One example of such a guarantee could consist of providing code to First Proof for setup in a secure cloud computing environment, allowing calls only to public models by independent providers.

Please contact us at contact@1stproof.org by **April 14, 2026** if you would be interested in such an assessment.

4 Benchmark: Grading (June, 2026)

The solutions produced by AI systems will be graded by human mathematicians in a manner similar to the review process in mathematics journals. The editorial board may decide, upon receiving the solution, to immediately reject it if the main statement it claims to prove is false. We will identify referees and provide them with anonymized author and AI solutions. The referees will be paid for each solution that they grade, and will be asked to gauge the

correctness of the solution (major errors, minor errors), and to assess the novelty of the solution (new ideas versus standard techniques).

The editorial board will obtain at least two reports for each submission, but may request more as required (e.g. if there is disagreement). The editorial board will decide if each solution is:

- essentially flawless (all disagreements are stylistic);
- publishable with minor revisions (i.e. a few minor errors which are easily corrected);
- requires major revisions (i.e. there is a flaw in the implementation of the strategy which require significant work to address);
- should be rejected (i.e. follows a strategy which is not clearly salvageable).

The editorial board will then 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 AI solutions, the referee reports, and the editorial board's decision along with its reasoning.

5 Community Experimentation (June, 2026)

For the informal community experimentation, we will make questions available to the public after we release the results of the formal benchmark, on a date announced **at least two weeks in advance** on our website. After a few days, we will post the solutions, which will be followed by a Zulip discussion. We will not formally verify autonomy or correctness, but request that participants include the following with their solutions, which will make it easier to interpret the results during the discussion : (1) a complete transcript of the interaction with AI models, fully specifying the details of human involvement, and (2) a record of the resources used, such as number or cost of tokens.

6 Funding

We will seek unrestricted donations from commercial companies whose systems we are testing. These donations will be allocated towards the costs required to create the questions and assess the solutions. Moreover, these donations will not be used to compensate editors or members of the board of directors. We will seek additional funding to test submissions by non-commercial entities. We are in the process of registering as a 501(c)(3) corporation and will release full reports of our expenses.

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.