IMProofBench - an AI benchmark for mathematical reasoning

Johannes Schmitt

Core objectives

- test AI ability to solve **problems from real-world mathematical research**, generating long-form arguments that **maintain rigor and avoid hallucinations**
- score AI output by both **human and AI graders** using a structured evaluation scheme; compare results to obtain a **dual benchmark for proof generation and evaluation**
- complement main question with **unique-answer subquestions** for fast automated grading
- curate and maintain the repository of problems within academia, with the majority kept fully private to **prevent training data contamination and model optimization against the benchmark**

Rationale for proof-based approach compared to alternatives

Some of the central problems of using current LLMs for math research are incorrect logical reasoning and hallucinated results. Benchmark problems that solicit long-form proofs directly target such behavior, effectively detecting and penalizing it in the evaluation. Moreover, by enabling the inclusion of more open-ended problems, this benchmark would fill a gap in the existing test frameworks, which focus on

- **Unique short-form answers (e.g. numerical results)** [FM24; HLE25]
  - not representative of mathematical research practice,
  - biased against abstract areas (like set theory, logic, ...) that often don't naturally yield numerical answers,
  - prone to short-cuts or educated guesses that unexpectedly reduce question difficulty.
- **Formalized proofs (e.g. in Lean)** [MiniF2F22]
  - restricted to mathematics that can be reasonably formalized, excluding much of cutting-edge research,
  - output format is underrepresented in current training corpus.

See also Appendix B for a comparison with other benchmarking approaches.

Overview

Below we give an overview of the planned benchmark structure, including the problem design and evaluation framework. We outline the associated organizational plans, from planned outreach activities and contributor recruitment to our motivations and measures to preserve the privacy of the benchmark. We also address major challenges such as the required scale of human grading and the acquisition of possible funding sources. Finally, we outline our plans for an initial pilot study for testing and refining the concept of the benchmark at smaller scale.

1 Benchmark structure

1.1 Scope of problems

1.1.1 Topics

Given the goal of testing AI reasoning in mathematics, the core subjects of benchmark questions will be in **Pure Mathematics**, with areas such as Algebra, Combinatorics, Geometry, Logic, Number Theory, etc. However, we will also reach out to other communities such as **Applied Mathematics**, **Theoretical Computer Science** and **Mathematical Physics** to solicit benchmark questions, given the importance of rigorous proofs in these areas.

1.1.2 Level of difficulty

Both for required background knowledge and complexity of the expected reasoning requirements, we focus on problems that are representative of those encountered during academic work at **PhD level and above**.[1] While we might do some exploratory trials with lower-level tasks (e.g. questions suited for an advanced undergraduate exam), we expect that the most recent reasoning models (like o3, o4-mini by OpenAI or Gemini 2.5 Pro by Google DeepMind) will show saturated performance on such problems.

1.2 Problem format

Each benchmark problem consists of

(a) a **main question**, whose answer is expected to be fully-justified by means of a (LaTeX-formatted, human-readable) proof,

(b) a collection of related **sub-questions** with unique answers (e.g. "Yes"/"No", a rational number, or a formal expression like $\sin(x)^2 \cdot \sqrt{\tan(y)}$) which can be checked by a simple Python script.

Ideally, the tasks in (b) are chosen such that a full solution of the main question is both necessary and sufficient to get a perfect score on these sub-questions. They serve as a cheap-to-evaluate and objective measure of the AI performance for the problem.

In addition to the question statements above, the full problem specification also contains

(c) detailed **sample solutions** for both main and sub-questions,

(d) a robust **grading scheme** for the main question, following academic standards (partial credits, penalties for particular mistake types) for distributing a total of 100 points per question,

(e) various types of **metadata**, such as mathematical area and estimated difficulty along different dimensions (background knowledge, complexity of reasoning, magnitude of required insights, or compute requirements).

See Appendix A for a collection of example problems, illustrating different types of questions suitable for the benchmark.

[1] In the framework of the FrontierMath benchmark [FM24] this would correspond to Tier 2 (graduate), Tier 3 (post-graduate) and – ideally – Tier 4 (expert research group).

1.3 Evaluation framework

Given the complexity of the questions, we plan to test the AI models in a multi-turn prompting environment with tool access, similar to the one established in Frontier Math [FM24]. Compared to the setup there, and in contrast to other math benchmarks, we will explore offering an expanded toolset, including

- **advanced computer algebra systems** like SageMath instead of simple Python environments,
- **online search and lookup** for reference papers and mathematical databases like the Online Encyclopedia of Integer Sequences.

Making such tools available offers a more realistic simulation of a modern mathematical research environment and removes artificial constraints (e.g. being unable to look up precise wordings of theorems or numerical data).

2 Organizational considerations

2.1 Outreach and institutional support

In the initial phase leading up to our smaller-scale pilot project we will focus on leveraging our personal mathematical networks to recruit contributors, likely focusing on the research area of the PI.

- **Ambassadors**: find highly motivated and well-connected individuals, who both contribute themselves and engage their colleagues and collaborators
- **Outreach at conferences**: offer short introductory presentation and informal question brainstorming sessions at research meetings (e.g. Oberwolfach workshop "Recent Trends in Algebraic Geometry" (June 2025) or the Summer Research Institute in Algebraic Geometry (July 2025))
- **Domain experts** (optional): recruit collaborators with complementary expertise to the main PI to help with building the benchmark, e.g.
  - *computer science*: experience with running AI benchmarks, web design or IT security to help with establishing necessary infrastructure
  - *statistics*: help with evaluating the benchmark results (e.g. which question features predict AI performance, comparison of human and AI grading), work out pre-registration for larger-scale study

Once some basic infrastructure and a proof-of-concept has been established via our pilot, we plan to expand the outreach to access a wider audience and gain credibility via:

- **Domain editors**: recruit established community members from different areas of mathematics and beyond, responsible for recruiting question authors and reviewers in their own field, to ensure balanced coverage
- **Advisory Board**: assemble a panel of respected mathematicians to oversee benchmark and provide credible testimonials on benchmark quality

- **Institutional partners:**
  - **ETH Zurich**: use existing IT infrastructure for publishing submission website, explore possibility of local hosting of open-source models (like DeepSeek R1) for question feedback or grading, to ensure question privacy
  - **MFO Oberwolfach**: Utilize their high throughput of research mathematicians as contributors, submit proposal for mini-workshop for developing benchmark infrastructure and focused problem creation
  - **Simons Center for Geometry and Physics**: Leverage existing connections and expertise, possibility to apply for a targeted grant of the Simons Foundation

2.2 Contributor motivation

- **Website**: build a secure, user-friendly project website for submitting and reviewing questions and grading AI answers, inspired by the corresponding website for Humanity's Last Exam [HLE25]; planned features include:
  - **Access to frontier models**: while working on questions, these can be tested against the latest AI models, offering a convenient way to experiment with these systems and query multiple AIs simultaneously
  - **Karma System**: reward points for completing tasks like question submissions, reviewing other questions, or grading AI solutions to incentivize participation and contributor retention; inspired by the corresponding system on the Stack Exchange websites
- **Co-authorship for all contributors** with author order possibly tied to the Karma system; depending on the scale of the project, multiple different publications on different aspects of the benchmark are possible (analysis of grading system, influence of prompt engineering and available toolset, qualitative analysis of common AI mistakes, etc)
- **Generous question retraction policies**: If authors discover that ideas from a question lead to research insights they would want to publish, they can request removal of the question from the benchmark. With their consent, these retracted problems could then be published among the public sample question set (making the benchmark quality more transparent while maintaining integrity). References to their published work would be included in the academic articles on the benchmark.

2.3 Quality assurance

- **Peer review**: benchmark organizers (or domain editors) solicit reviews of submitted questions, either via the website or by reaching out to experts, similar to the practice within academic publishing. Compared to a full referee report of a journal article, these solicitations are much more self-contained and less time-consuming, and might double as a way to further spread the word about the benchmark
- **Competitions for human baseline**: test problems with human mathematicians to calibrate difficulty (e.g. via local events organized by benchmark ambassadors)
- **Transparency**: all methodological aspects except the actual problems are publicly documented (e.g. via a GitLab repository) and open to standard open development procedures like Pull Requests, Issues etc.

- **Evaluation via grading**: typical error rates in the required answers for Machine Learning benchmarks can range from 5 - 10 % (see e.g. [Gem+25]). With human (in-the-loop) grading, it is much more likely to detect such issues once the AI finds the correct solution (e.g. by convincing the grader of the corrected argument). Additionally, we will explore options for the AI graders to flag potential issues with solutions to the benchmark organizers.

2.4 Grading at scale

Reading and evaluating mathematical proofs is quite time-consuming; with many models to test, a thorough human evaluation of the answers to the main questions can quickly become challenging. The design of the benchmark incorporates multiple measures to address or mitigate this issue. Firstly, during the initial pilot phase of the project, we plan to gather data on the grading process:

- **Tracking of human effort**: The grading time of human evaluators (initially: likely the author of the question) will be recorded. The results might indicate that the total effort is manageable, in particular for the initial grading when the answers of all existing models are graded in a row, while the question and grading scheme is still fresh on the mind of the author.
- **Autonomous AI grading**: We plan to compare the reliability of AI grading versus the human expert grading, given full solutions and evaluation scheme. Depending on the quality of the AI grading, it might be feasible to use it as the main scoring system for benchmark runs, with only sporadic human grading as validation.
- **AI-Assisted Grading**: One hybrid option is to use AI models to *assist* human graders (e.g. making an initial proposal, flagging critical parts of the proposed solution or summarizing parts of the argument). This could reduce the required human effort, and is also of independent scientific interest, e.g. due to possible applications to other settings (e.g. exam grading assistance in university teaching).

In addition to the strategies explained above, which depend on the outcome of the pilot study, there are further general options to handle the grading:

- **Grader recruitment**: With detailed sample solutions (especially for lower-tier difficulty), it could be feasible to involve people other than the original author and reviewer. Possible rewards are participation credits (from the Karma system mentioned above) and associated co-authorship, and even monetary compensation (as in the commercial Outlier AI platform) if sponsors or other funding sources can be obtained.
- **Unique-answer sub-questions as fallback** Even if the grading of the main question remains infeasible, the benchmark has the automated evaluation of the included sub-questions as a cheap and quick alternative.

2.5 Benchmark privacy

ProofBench is planned as a benchmark with a majority closed dataset:

- **Public sample questions**: small, representative public set of problems, to illustrate typical structure and difficulty
- **Fully private evaluation set**: majority of questions remains confidential, with measures to ensure secrecy (see below)

The main arguments for this design decision are:

- **Prevent model optimization on the benchmark** One phrasing of Goodhart's law states

  > When a measure becomes a target, it ceases to be a good measure. [Str97]

  With open benchmark datasets, model developers can either train on the test-set outright or perform A/B testing by running the benchmark frequently, to guide high-level design decisions (as was the case for FrontierMath with OpenAI). Both of these practices erode the reliability of and trust in the original benchmark. With a fully private test set we aim to preserve the integrity of our measurement and provide unbiased evaluations.

- **Avoid contributing to AI capabilities research** The focus on measuring progress without providing possible training data and guidance is an important consideration for potential contributors who are wary of accelerating the development of AI capabilities.

Planned measures for ensuring that the evaluation set remains confidential:

- **Existing zero data retention policies** All major frontier AI labs already offer policies for keeping customer data private and many have zero retention policies for API calls, ensuring that sent data is deleted once processing has been completed. We plan to reach out to all AI labs whose models will be tested to ensure that information transmitted for evaluating or AI-grading model answers is not stored or used by the company, and might seek additional legal assurances.

- **Legal framework**: We plan to draft robust and transparent (i.e. publically available) agreements with contributors and AI companies to ensure that confidentiality of the benchmark problems is maintained.

- **Local model evaluation**: Whenever feasible, we plan to use locally-hosted open-source models for initial screening, question feedback or AI grading.

- **Information compartmentalization**: Beyond a certain scope of the project, we take measures to ensure individual contributors only see limited subsets of problems.

- **Technical verification**: We plan to employ watermarking or fingerprinting techniques to detect if problems appear in training data (e.g. including the BIG-bench canary string in our problem dataset as a basic precaution measure).

- **Benchmark structure**: With human evaluations being a central component of the benchmark, it is not possible to run it without expert human help, even if given full data access.

2.6 Pilot project

To validate and refine the feasibility of our benchmark design, and build the required infrastructure, we plan to have an initial smaller-scale trial:

- **Scope**: 25-50 questions, possibly focused on algebraic geometry (to profit from the professional network and expertise of the PI),

- **Timeline**:
  - May-June 2025: collecting first benchmark problems, implementing basic evaluation infrastructure
  - July-August 2025: first larger-scale AI evaluations with response grading
  - Sept 2025: finalizing the experimental stage, write-up of results (e.g. for submission of a conference paper to ICLR 2026)
- **Reduced complexity**: manual submission, review and grading instead of polished contributor website, less strict IT and information security around benchmark questions, private funding of API calls

2.7 Funding

Apart from the institutional partnerships mentioned before, there are several other sources of (financial) support, in particular once some basic credibility has been established via a pilot study.

- **Academic research grants**: We will reach out both to existing research groups and projects, whose scope covers the planned benchmark, and write grant proposals (e.g. for project funding of the Swiss National Science Foundation, to cover personnel or API costs).
- **AI research labs** Since an independent evaluation of their models is broadly in the interest of leading AI labs (both for scientific and PR reasons), we will reach out to them to propose collaborations and support from the labs (e.g. via donations of API credits). While we are committed to the benchmark privacy as explained above, we could offer more detailed qualitative and quantitative analysis of strengths and weaknesses of their respective models in exchange.
- **Private philanthropy**: In previous years there were examples of significant financial support for similar projects by charitable individuals and foundations (see e.g. the sponsor list of EpochAI and the AI for Math Fund). We believe that our proposal offers a best-in-class approach for AI evals from the perspective of benchmark integrity and avoiding capability acceleration, and would be in an excellent position to apply for such future funding rounds.

References

[Gem+25] A. P. Gema et al. *Are We Done with MMLU?* 2025. arXiv: 2406.04127 [cs.CL].

[FM24] E. Glazer et al. *FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI.* 2024. arXiv: 2411.04872 [cs.AI].

[USAMO25] I. Petrov et al. *Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad.* 2025. arXiv: 2503.21934 [cs.CL].

[HLE25] L. Phan et al. *Humanity's Last Exam.* 2025. arXiv: 2501.14249 [cs.LG].

[Str97] M. Strathern. "'Improving ratings': audit in the British University system". In: *European Review* 5.3 (1997), pp. 305–321. DOI: 10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

[MiniF2F22] K. Zheng, J. M. Han, and S. Polu. *MiniF2F: a cross-system benchmark for formal Olympiad-level mathematics.* 2022. arXiv: 2109.00110 [cs.AI].

A Example problems

Below we give a few samples of (types of) problems that are suitable for the benchmark, edited for brevity (full specification would include more details on the definitions etc).

A.1 Complex formula answer, subquestions check evaluations

Main question: Find a concise expression for the number $N(g)$ of stable graphs of genus g with no legs and precisely 3 edges, for all $g \geq 2$.
Solution: $N(g)$ is a quasi-polynomial with period 6 for $g \geq 3$, calculated by summing contributions from each topological type of stable graph
Subquestions: What are the values of $N(g)$ for $g = 3, 8, 1000$?

A.2 Proof-or-counterexample for claim, subquestions check validity of variations

Main questions: For a finite group G, is the functor Perm : G-sets $\to \mathrm{Rep}_{\mathbb{C}}(G)$ sending X to the complex permutation representation $G \curvearrowright X^{\mathbb{C}}$ well-defined and fully faithful?
Solution: The functor is not fully faithful, a concrete counter-example (of non-isomorphic G-sets mapping to the same permutation representation) can be constructed for $G = (\mathbb{Z}/2\mathbb{Z})^3$.
Subquestions: Is the claim from the main question true for a) all finite groups G, b) all finite abelian groups, c) all finite cyclic groups, d) all compact Lie groups G acting [...]?

A.3 Open-ended classification problem, subquestions check numerical consequences

Main questions: Let $X = \overline{\mathcal{H}}_{g\to 0,d}$ be the admissible cover compactification of the Hurwitz space of degree d covers from a genus g curve to $\mathbb{P}^1$. Give a combinatorial classification of the cones and face-morphisms in the tropicalization Σ_X.
Solution: Cones of Σ_X correspond to covers of stable graphs, with additional ramification and monodromy representation data at the vertices, face-morphisms correspond to edge-contractions.
Subquestions: What is the number of isomorphism classes of cones of Σ_X for $(g, d) = (3, 2), (5, 2), (6, 10)$?

B Comparison with related benchmarks

The benchmark addresses gaps in existing frameworks like:

- **MATH/MATH2**: Focuses on high school competition problems, now nearly saturated
- **Omni-MATH/OlympiadBench**: Olympiad-level but not research-level mathematics
- **Putnam-AXIOM**: Undergraduate competition level, below research mathematics
- **Math Arena - USAMO** [USAMO25]: Human grading of AI answers to US Math Olympiad problems
- **FrontierMath** [FM24]: True research-level but focuses on unique answers, which can sometimes be reached through shortcuts or guessing without thorough understanding
- **miniF2F** [MiniF2F22]: Problems are easier than research-level, already public, and focused on formal verification rather than natural language proofs

This proposal combines aspects of these benchmarks while emphasizing proof generation and evaluation at the research level, with strong privacy preservation and a focus on authentic mathematical practice.