Can a Simple Automated AI Pipeline Solve Research-Level Mathematical Problems?

AI Mathematics Research Team

February 11, 2026

Abstract

Large language models (LLMs) have recently achieved remarkable success in generating rigorous mathematical proofs, with AI for Math emerging as a vibrant field of research [5]. While these models have mastered competition-level benchmarks like the International Mathematical Olympiad [3, 4] and show promise in research applications through autoformalization [9], their deployment via lightweight, natural-language pipelines for *de novo* research problems remains underexplored. In this work, we demonstrate that next-generation models (e.g., Gemini 3 Pro, GPT-5.2 Pro), when integrated into a streamlined automated pipeline optimized for **citation verification**, can solve sophisticated research-grade problems. We evaluate our pipeline on two novel datasets: (1) the **ICCM problem sets** (of Yau competition level) proposed by leading mathematicians [7], and (2) the **First Proof problem set** [1], consisting of previously unpublished research questions. Our pipeline successfully solved all problems in the first two ICCM sets, and claimed solutions for all 10 problems in the First Proof set; we have rigorously verified the solution to Problem 4, confirming its correctness. These results suggest that 2026 marks a turning point where AI becomes a practical assistant for genuine mathematical research. All generated proofs for the ICCM sets have been submitted to the official organization, and our evaluation results are publicly available at `https://your-link-here.com`. We plan to open-source the complete pipeline methodology subsequently.

1 Introduction

Automating advanced mathematical reasoning, a long-standing goal in AI, has accelerated markedly with the development of large language models (LLMs). Models have rapidly progressed from solving grade-school word problems to achieving medal-level performance in prestigious competitions [2][10], such as the International Mathematical Olympiad (IMO) [3]. This progression demonstrates the dual role of AI in mathematics: it provides robust tools for mathematical research while simultaneously serving as a premier testbed for advancing general reasoning in AI systems [5]. A question appears naturally: does success in solving curated contest problems translate to the ability to assist with genuine mathematical research?

Research-level mathematics is qualitatively different. As noted by Fields Medalist Shing-Tung Yau, while AI excels at high-dimensional computation, mathematicians remain essential for tackling deep, long-standing problems, and the future lies in collaboration [7]. The core of research often lies not in answering well-posed questions but in formulating them and developing

new frameworks [1]. Current benchmarks, primarily based on competition problems, fail to capture this reality. A critical step forward is the introduction of problem set like First Proof [1], which consists of previously unpublished, research-originated questions. This addresses a key limitation: data contamination, where a model's performance is inflated by having seen similar problems during training. As it is built from problems sourced directly from unpublished research, the design of the First Proof benchmark guarantees that solving them demands novel reasoning, moving beyond pattern matching.

Meanwhile, the mathematical community has begun to issue direct challenges. Recently, leading mathematicians at the International Congress of Chinese Mathematicians (ICCM) publicly posed a set of sophisticated problems, explicitly targeting AI systems to probe the boundaries of human knowledge [7]. These parallel developmentsthe creation of pristine research-level benchmarks and the public issuance of expert challengescreate a timely test method for evaluating AI's true research potential.

Bridging the gap between contest performance and research utility remains non-trivial. Auto-formalization methods, which translate statements into verifiable code (e.g., Lean), offer guaranteed correctness [9] but impose a high technical barrier, limiting accessibility for most mathematicians. We argue that a complementary path is crucial: developing lightweight, natural-language pipelines that can generate reliable, human-readable proofs. In this work, we demonstrate that by integrating next-generation LLMs (e.g., Gemini 3 Pro, GPT-5.2 Pro) into a streamlined automated pipeline optimized with a **citation verification** mechanism, we can reliably solve sophisticated research-grade problems. Our pipeline mandates explicit bibliographic grounding for non-trivial claims, significantly enhancing verifiability. We test this system on the two novel and challenging datasets described above: (1) the ICCM problem sets (of Yau competition level) [7], and (2) the First Proof problem set [1]. Our pipeline successfully solved all problems in the first two ICCM sets and claimed solutions for all ten problems in the First Proof set. We have rigorously verified the solution to Problem 4 of the latter, confirming its mathematical correctness.

These results suggest a turning point. By 2026, AI, when equipped with the right methodological scaffolding, appears capable of transitioning from a tool for solving known puzzles to a practical assistant for tackling components of genuine mathematical research.

2 Methodology

2.1 The Pipeline

We adopted the automated pipeline architecture proposed in [3], originally designed for IMO-level problems. Given the increased complexity of our target tasks (Yau competition and research level), we introduced two key modifications:

1. **Domain-Specific Prompt Optimization:** We refined prompts to handle higher-order abstract reasoning, moving beyond high-school olympiad strategies to incorporate undergraduate and graduate-level conceptual frameworks.

2. **Citation-Augmented Verification:** A critical limitation of previous pipelines was the

hallucination of theorems or formulas without context, rendering proofs unverifiable. To address this, we enforced a strict constraint: the model must provide specific bibliographic references for non-trivial claims and explain the role of each cited source in the argument.

2.2 Validation of the Method

To validate the citation-augmented approach, we tested the pipeline on representative exercises from Kashiwara's classic text *Categories and Sheaves* [6]. To probe the capability boundaries of the AI, several problems were selected by an independent researcher who had manually solved all the exercises in this book. The AI not only produced correct proofs but also correctly cited specific sections of the book. This significantly improved the interpretability of the output, allowing readers—even those less familiar with the text—to verify the logical chain.

3 Experiments and Results

We test our pipeline on two primary sources comprising extremely challenging mathematical problems.

3.1 ICCM Problem Sets

We utilized the three sets of problems proposed by the International Congress of Chinese Mathematicians (ICCM). Each set comprises six problems spanning several prominent mathematical domains, such as combinatorics, algebra, analysis, and differential equations.

3.1.1 Sets 1 & 2

These correspond to the difficulty of the S.-T. Yau College Student Mathematics Contest.

Result: Our pipeline successfully solved 100% of the problems in these two sets. The difficulty of these problems corresponds to the S.-T. Yau College Student Mathematics Contest level. The verification of the AI-generated solutions was firstly conducted by our team, which includes members with a background in pure mathematics and a recipient of the contest's all-round medal. The final, verified proofs have been compiled into a PDF and submitted to the ICCM organization.

3.1.2 Set 3 (Conjecture Level)

This set contains open problems. Section 1 includes famous conjectures unsolved for decades; Section 2 includes open problems related to Calabi-Yau manifolds.

Result: The AI failed to solve Section 1 (as expected for open conjectures). Section 2 was attempted but remains unverified due to the lack of specialized domain experts in our team.

3.2 The "First Proof" Problem Set

We conducted a complete test using the First Proof problem set [1], which consists of ten previously unpublished research-level questions originating from mathematicians ongoing work.

Following the benchmark protocol, all questions were tested on February 9, 2026, prior to the release of the official answers on February 13, 2026.

Result: Our pipeline claimed to have produced correct solutions for all ten problems. Given the complexity of the generated proofs and time constraints for human verification, we prioritized a thorough verification of Problem 4. A team of experts with doctoral degrees in pure mathematics confirmed that the AIs solution is mathematically sound and correct. Considering the pipelines demonstrated tendency to acknowledge its limits on genuinely intractable tasks (e.g., the open conjectures in ICCM Set 3, Section 1), its confident solutions across the entire First Proof set suggest a high probability of success for the remaining unverified problems.

4 Discussion

Our results indicate that the combination of simple automated pipelines and state-of-the-art LLMs has crossed a significant threshold, demonstrating a tangible capacity for enough mathematical reasoning on research-grade problems. However, this technical advancement reveals a shifting bottleneck and several practical hurdles that must be addressed to integrate AI as a practical assistant in daily mathematical research.

4.1 The Verification Bottleneck

The primary challenge has shifted from **proof generation** to **efficient verification**. While our pipeline produced candidate solutions for the entire First Proof set in minutes, the rigorous verification of a single problem (Problem 4) required hours. This asymmetry underscores an urgent need for more sophisticated AI-assisted verification toolswhether through advances in formal methods, interactive semi-formal interfaces, or explainable reasoning frameworksto keep pace with the accelerating speed of AI-generated mathematics.

4.2 Practical Hurdles for Applications In large scale

Beyond verification, our testing and development process revealed several significant challenges in applying AI systems to authentic mathematical research:

1. **Usability and Accessibility Gap**: Many practicing mathematicians are unfamiliar with the prompting techniques and effective use of advanced AI systems. Therefore, lowering the technical barrier and developing intuitive, powerful tools are crucial for widespread adoption within the mathematical community.

2. **Long-Context Reasoning**: Genuine research often involves grappling with long, interconnected chains of reasoning and multiple related sub-problems. As problems deepen, the required context length and the need for coherent long-term memory pose challenges to current AI architectures, potentially leading to fragmented or inconsistent reasoning.

3. **Understanding Implicit Knowledge**: Mathematical literature frequently contains implicit steps, assumed background knowledge, and notational shortcuts. When an AI lacks deep, specific domain understanding, it can fail to comprehend these jumps, as evidenced

by the observation that training on large corpora of raw arXiv papers alone (e.g., in early iterations like DeepSeekMath-V1 [8]) did not yield expected gains in deep comprehension. This highlights that merely scaling data is insufficient. A promising direction may involve employing AI to process mathematical literature at scale, explicitly reconstructing intermediate steps and completing logical chains, and subsequently using this augmented corpus for reinforcement learning or fine-tuning training.

4.3 Outlook and Future Work

Despite these challenges, the field of AI for Math shows immense promise. The successes documented in this work suggest that 2026 may indeed be a turning point. The future likely lies in collaborative synergy: AI can handle computationally intensive exploration, pattern suggestion, and tedious verification of sub-steps, freeing mathematicians to focus on high-level conceptualization and creative problem formulation. To realize this vision, future work need to focus on creating more intuitive interfaces, developing models with enhanced reasoning coherence, and fostering a deeper comprehension of mathematical literature.

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, 2026. URL https://arxiv.org/abs/2602.05192.

[2] Jiangjie Chen, Wenxiang Chen, Jiacheng Du, Jinyi Hu, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Wenlei Shi, Zhihong Wang, Mingxuan Wang, Chenrui Wei, Shufa Wei, Huajian Xin, Fan Yang, Weihao Gao, Zheng Yuan, Tianyang Zhan, Zeyu Zheng, Tianxi Zhou, and Thomas Hanwen Zhu. Seed-prover 1.5: Mastering undergraduate-level theorem proving via learning from experience.

[3] Yichen Huang and Lin F. Yang. Winning gold at imo 2025 with a model-agnostic verification-and-refinement pipeline, 2025. URL https://arxiv.org/abs/2507.15855.

[4] IMO Official. International mathematical olympiad, 2025. URL https://www.imo-official.org. Accessed: 2026-02-11.

[5] Haocheng Ju and Bin Dong. Ai for mathematics: Progress, challenges, and prospects, 2026. URL https://arxiv.org/abs/2601.13209.

[6] Masaki Kashiwara and Pierre Schapira. *Categories and Sheaves*, volume 332 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2006. ISBN 978-3-540-27949-5. doi: 10.1007/3-540-27950-4.

[7] Shanghai Institute of Mathematical Sciences and Interdisciplinary Sciences. Shanghai issues the "mathematical questions" again after six months: Three challenging math problems pushing ai to its limits. DongJing (WeChat Official Account Platform), 1 2026. URL https://mp.weixin.qq.com/s/v2KA8Cdoe-Nj0599GIfQMw. A non-peer-reviewed news report. It

covers the event where Fields Medalist Shing-Tung Yau and other mathematicians publicly issued challenging math problems to AI systems. The article also mentions the performance of several AI models, including InternLM, ByteDance Seed, Qwen, and SenseTime, in tackling these problems. Suitable for citation as background information or to reference a current development.

[8] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[9] Hanyu Wang, Ruohan Xie, Yutong Wang, Guoxiong Gao, Xintao Yu, and Bin Dong. Aria: An agent for retrieval and iterative auto-formalization via dependency graph, 2025. URL https://arxiv.org/abs/2510.04520.

[10] Chengda Lu Z.Z. Ren Jiewen Hu Tian Ye Zhibin Gou Shirong Ma Xiaokang Zhang Zhihong Shao, Yuxiang Luo. Deepseekmath-v2: Towards self-verifiable mathematical reasoning, 2025.