Winning Gold at IMO 2025 with a Model-Agnostic Verification-and-Refinement Pipeline

Yichen Huang (黄溢辰)
huangtbcmh@gmail.com

Lin F. Yang (杨林)
University of California, Los Angeles
linyang@ee.ucla.edu

Abstract

The International Mathematical Olympiad (IMO) is widely regarded as the world championship of high-school mathematics. IMO problems are renowned for their difficulty and novelty, demanding deep insight, creativity, and rigor. Although large language models perform well on many mathematical benchmarks, they often struggle with Olympiad-level problems. Using carefully designed prompts, we construct a model-agnostic, verification-and-refinement pipeline. We demonstrate its effectiveness on the recent IMO 2025, avoiding data contamination for models released before the competition. Equipped with any of the three leading models—Gemini 2.5 Pro, Grok-4, or GPT-5—our pipeline correctly solved 5 out of the 6 problems ($\approx 85.7\%$ accuracy). This is in sharp contrast to their baseline accuracies: 31.6% (Gemini 2.5 Pro), 21.4% (Grok-4), and 38.1% (GPT-5), obtained by selecting the best of 32 candidate solutions. The substantial improvement underscores that the path to advanced AI reasoning requires not only developing more powerful base models but also designing effective methodologies to harness their full potential for complex tasks. Code available at: `https://github.com/lyang36/IMO25`

1 Introduction

The International Mathematical Olympiad (IMO) is an annual competition widely regarded as the world championship of high-school mathematics. Established in Romania in 1959 with just seven participating countries, it has since expanded to include over 100 nations, each represented by a team of up to six of their most talented pre-university students. The competition is an extremely challenging test of creativity and sustained concentration: over two consecutive days, contestants are given two 4.5-hour sessions to solve three problems per session, drawn from the fields of algebra, combinatorics, geometry, and number theory [2].

Qualifying for the IMO is itself extremely challenging. In the United States, for instance, a student must advance through a rigorous series of national competitions of increasing difficulty, from the American Mathematics Competitions (AMC) to the American Invitational Mathematics Examination (AIME), and finally to the USA Mathematical Olympiad (USAMO). Top performers at USAMO are invited to compete for the six spots on the U.S. national team. Most other countries have similarly stringent selection processes, ensuring that the IMO convenes the world's most talented pre-university students. A gold medal at the IMO is an extraordinary achievement, awarded to only the top twelfth of the contestants. Consequently, the IMO serves as the preeminent stage where future leaders in mathematics demonstrate their exceptional talent, and success at the IMO has a significant correlation with the Fields Medal, the highest honor in mathematics. Of the 34 Fields medalists awarded since 1990, 11—including renowned mathematicians Terence Tao, Maryam Mirzakhani, and Grigori Perelman—are prior IMO gold medalists.[1] Furthermore, the probability that an IMO gold medalist

[1] Compiled by the authors by checking the IMO records of all Fields medalists awarded 1990-2022. IMO data was sourced from the official website.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

will become a Fields medalist is 50 times larger than the corresponding probability for a PhD graduate from a top-10 mathematics program [6].

Advanced mathematical reasoning is a hallmark of intelligence and the foundation of science and technology. Consequently, automated mathematical reasoning has become a major frontier in artificial intelligence (AI). The rapid advancement of Large Language Models (LLMs) has enabled them to master mathematical benchmarks of increasing difficulty [7, 29]. This progress has been enabled by inference-time methods such as Chain-of-Thought, which improves performance on complex tasks by breaking them down into a sequence of intermediate reasoning steps [31]. Early datasets GSM8K [13] and MATH [18], which test grade-school and high-school mathematics, respectively, have been largely solved. The performance of leading models, such as Gemini 2.5 Pro [16], Grok-4, and GPT-5, is also approaching saturation on the AIME, a significantly more challenging competition benchmark. However, AIME problems are not required to be entirely novel: 8 out of the 30 problems in AIME 2025 were identified as having close analogs in online sources available prior to the event [9]. This allows models to achieve high performance partly through sophisticated pattern recognition and adaptation of existing solutions rather than completely original reasoning.

The remarkable success of LLMs on these benchmarks has pushed the frontier of AI mathematical reasoning to the next tier: Olympiad-level problems [17]. This represents a shift not merely in difficulty, but in the very nature of the task. Whereas the AIME requires only a final numerical answer, the USAMO and IMO demand a complete and rigorous proof. In mathematics, an answer without a rigorous proof is merely a conjecture; it is the proof that promotes a conjecture to a theorem. Furthermore, IMO problems are systematically selected for novelty: the selection process is designed to filter out any candidate problem that is too similar to a known problem [4, 20]. Thus, solving IMO problems requires original insights and multi-step creative reasoning, rather than pattern recognition and retrieval from training data. These three pillars—the renowned difficulty, the demand for rigor, and the strict criterion of problem novelty—establish the IMO as a grand challenge and the preeminent benchmark for assessing the genuine mathematical reasoning capability of LLMs. The demand for logically sound arguments, in particular, exposes a critical weakness in current LLMs [24]. Recent evaluations on the USAMO 2025 [26] and IMO 2025 [3] show that state-of-the-art models struggle to generate sound, rigorous proofs, often committing logical fallacies or using superficial heuristics, and consequently fail to win even a bronze medal.

Last year, Google DeepMind announced a breakthrough: an AI system that achieved a silver-medal performance at the IMO 2024 [8]. Their approach used AlphaGeometry 2 [27, 12], a specialized solver for geometry, and AlphaProof for algebra and number theory problems. Notably, AlphaProof generates proofs in the formal language Lean. The primary advantage of this formal approach is guaranteed correctness: a proof successfully verified by the Lean proof assistant is irrefutably sound. However, this guarantee comes at the cost of human readability. Proofs in formal languages are often verbose and cumbersome, and require specialized training to understand, making them inaccessible to most mathematicians. Our work, in contrast, is situated entirely within the natural language paradigm. Our approach produces human-readable proofs, akin to those in mathematical journals and textbooks. This is crucial for enabling effective human-AI collaboration, where mathematicians can understand, critique, and build upon an AI's reasoning. While generating machine-verifiable proofs is a vital goal, our natural language approach tackles the complementary challenge of creating an AI that can reason and communicate like a human mathematician, thereby making its insights easily accessible to the scientific community.

In this paper, we construct a model-agnostic verification-and-refinement pipeline and demonstrate its effectiveness across three leading large language models: Gemini 2.5 Pro, Grok-4, and GPT-5. When equipped with any of these models, our pipeline solved 5 out of the 6 problems from the IMO 2025, achieving an accuracy of approximately 85.7%. This result stands in sharp contrast to the models' baseline performance. An independent evaluation [3] by MathArena employed a best-of-32 post-selection strategy: for each problem, 32 solutions are generated, and the model itself selects the most promising one for human grading. Even with this performance-boosting inference-time method, the reported accuracies were only 31.6% for Gemini 2.5 Pro, 21.4% for Grok-4, and 38.1% for GPT-5. A pervasive and fundamental challenge in the evaluation of LLMs is data contamination, where test problems are included in a model's training data, inflating its performance metrics [11]. Our use of the recent IMO 2025 problems helps mitigate this issue. Since Gemini 2.5 Pro and Grok-4 were released before the competition, they were evaluated on a pristine testbed. While GPT-5 was released after the competition, raising the possibility of data contamination, the comparison to the best-of-32 baseline remains fair, as any contamination would affect both their and our evaluations. The substantial performance gain—from a baseline of 38.1% to our 85.7%—isolates the contribution of our pipeline, confirming its effectiveness regardless of potential data contamination. Our results demonstrate that strong existing LLMs already possess powerful mathematical reasoning capabilities, but that a verification-and-refinement pipeline is essential for converting their latent capabilities into rigorous mathematical proofs.

To further validate the generalizability and robustness of our pipeline, its performance was independently assessed on a different and challenging benchmark: the 2025 International Mathematics Competition for University Students (IMC). The IMC is a prestigious annual contest that includes topics from undergraduate curricula. Thus, it requires a broader and more advanced mathematical knowledge base than the IMO. MathArena evaluated what they termed the "Gemini agent"—our pipeline with Gemini 2.5 Pro as the base model. The agent achieved 94.5% accuracy [5], ranked #3 among 434 human participants. By contrast, the base model alone only scored 57.7% and ranked #92. This third-party validation demonstrates our pipeline's effectiveness in a more knowledge-intensive domain and on a pristine, uncontaminated dataset, as the competition occurred after the public release of our code.

Our work builds upon a growing body of research aimed at enhancing the reasoning capabilities of LLMs through verification and iterative refinement. Foundational works [21, 23] have pioneered the framework of this approach, where a model generates an output, receives feedback, and then refines its work. In the mathematical domain, this approach has been adapted into various methods designed to verify and improve the logical steps of a solution [32, 25, 28]. Another line of research focuses on generating and repairing proofs in formal languages to guarantee correctness [15]. While our pipeline is built on these core ideas of iterative refinement, our contribution is to construct a model-agnostic, inference-time framework with carefully designed prompts that specialize this process for the extreme rigor and novelty demanded by Olympiad mathematics. Our robust verifier design directly addresses the challenge of generating high-quality feedback, a known bottleneck for self-correction methods [19]. By applying our pipeline to state-of-the-art LLMs, we demonstrate a level of performance on the IMO 2025—a grand-challenge benchmark of significantly greater difficulty than those addressed in prior studies—that was previously unattainable.

Other teams, including OpenAI [30], Google DeepMind [22], and ByteDance [10], announced strong performance of their AI systems on the IMO 2025 problems after the event.

2 Pipeline

Overview. At a high level, our pipeline proceeds as follows (illustrated in Figure 1):

- Step 1: Initial solution generation with the prompt in Appendix A.1;
- Step 2: Self-improvement;
- Step 3: Verifying the solution with the prompt in Appendix A.2 and generating a bug report; go to Step 4 or Step 6 (see below for explanations);
- Step 4: Review of the bug report (optional);
- Step 5: Correcting or improving the solution based on the bug report; go to Step 3;
- Step 6: Accept or Reject.

We run the procedure some number of times (in parallel or in serial, independently) in order to obtain a correct solution. We hope that the model either outputs a correct solution or reports that it failed to find one.

Detailed Workflow. The solver prompt in Appendix A.1 for Step 1 is designed to emphasize rigor rather than focus on finding the final answer and thus matches the theme of IMO. We have randomly selected some outputs of this step and found that the overall quality of the solutions is pretty low. This is consistent with very recent findings of Ref. [3].

In Step 2, the model is prompted to review and try to improve its work. General-purpose LLMs are not tailored to solving exceptionally challenging mathematical problems in a single pass. A significant constraint is their finite reasoning budget allocated for a single query. For instance, the

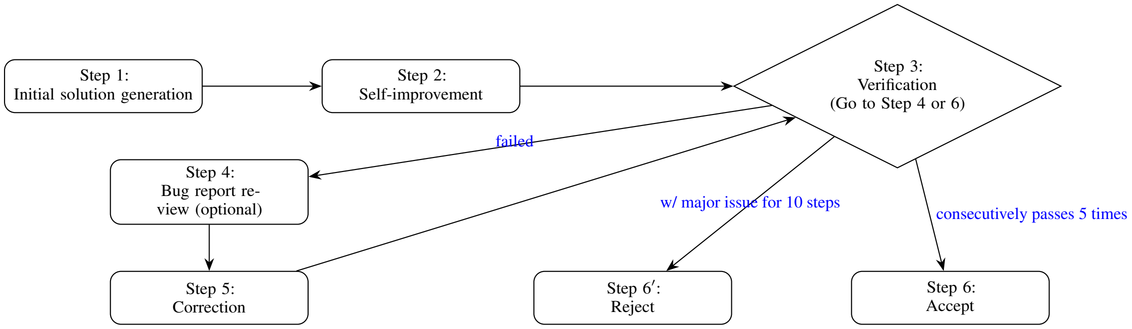

Figure 1: Flow diagram of our pipeline. See the main text for detailed explanations of each step.

maximum number of thinking tokens for Gemini 2.5 Pro is 32,768, which is not enough for solving a typical IMO problem. We observe that in Step 1, the model almost always uses up this budget. Consequently, it lacks the capacity to fully solve the problem in one go. This is why we break down the problem-solving process into steps. Step 2 effectively injects another budget of reasoning tokens, allowing the model to review and continue its work. We consistently observe that the outputs have been noticeably improved during Step 2.

Next we will use the verifier to make iterative improvement and decide whether to accept an improved solution.

The verifier plays an important role in our pipeline. Its functionality is to carefully review a solution step by step and find out issues (if any). We emphasize mathematical rigor and classify issues into critical errors and justification gaps. Critical errors are something that is demonstratively false or with clear logical fallacies, while justification gaps can be major or minor. A major justification gap that cannot be repaired would crash an entire proof, while minor justification gaps may not even be well defined: A minor gap could sometimes be viewed as concise argument.

In Step 3, we use the verifier to generate a bug report for each solution outputted in Step 2. The bug report contains a list of issues classified as critical errors or justification gaps. For each issue, an explanation is required. The bug report will serve as useful information for the model to improve the solution, either fixing errors or filling gaps. Step 4 (optional) is to carefully review each issue in the bug report. If the verifier makes a mistake and reports an issue which is not really an issue, the issue would be deleted from the bug report. Thus, Step 4 increases the reliability of the bug report. In Step 5, the model tries to improve the solution based on the bug report. We iterate Steps 3-5 a sufficient number of times until we decide to accept or decline a solution. We accept a solution if it robustly passes the verification process and decline a solution if there are always critical errors or major justification gaps during the iterations.

3 Results and Discussion

Performance on IMO 2025. Our model-agnostic pipeline demonstrated consistent success across three leading LLMs. When equipped with Gemini 2.5 Pro, Grok-4, or GPT-5, the pipeline successfully generated rigorous solutions for 5 out of the 6 problems from the IMO 2025. The full, verbatim proofs for each problem from each model, which constitute the primary evidence for this claim, are provided in Appendix B.

Despite the high success rate, the pipeline failed to solve Problem 6, and this failure was consistent across all three base models. The consistent failure on this problem suggests that certain types of complex combinatorial reasoning remain a significant hurdle for current models, even within a verification-and-refinement framework.

Recent findings by MathArena [3] highlight a key challenge: single-pass solution generation is often insufficient for complex tasks demanding mathematical rigor. This is evidenced by the baseline accuracies on the IMO 2025, where even a best-of-32 post-selection strategy yielded only 31.6% for Gemini 2.5 Pro, 21.4% for Grok-4, and 38.1% for GPT-5. In sharp contrast, our pipeline achieved a consistent accuracy of approximately 85.7% across all three models. This substantial improvement demonstrates that the iterative refinement process systematically overcomes the limitations of single-pass generation, such as finite reasoning budgets and the critical errors or justification gaps that often appear in initial drafts. The verifier-guided loop, in particular, proved essential for eliciting rigorous and trustworthy arguments, validating the central thesis of this work: such a pipeline is key to converting the latent capabilities of powerful LLMs into sound mathematical proofs.

Generalization to Undergraduate Mathematics. To assess the broader applicability and robustness of our pipeline, it is essential to evaluate its performance on benchmarks that differ significantly from the IMO. We therefore turn to the International Mathematics Competition for University Students (IMC), a prestigious annual contest held since 1994. Like the IMO, the IMC requires complete and rigorous proofs, making it an excellent testbed for our verification-focused methodology. The IMC problems are drawn from the fields of algebra, analysis (real and complex), geometry, and combinatorics, reflecting the core of a standard undergraduate mathematics curriculum. Thus, the IMC requires a more extensive and advanced knowledge base than the IMO, which is grounded in pre-university mathematics. Its long history and emphasis on rigorous proofs establish the IMC as a well-regarded benchmark for advanced mathematical reasoning.

MathArena independently evaluated our pipeline on the IMC 2025 [5]. They implemented our publicly available code with Gemini 2.5 Pro as the base model, referring to this implementation as the "Gemini agent." The results demonstrated a substantial performance improvement attributable to our pipeline. The Gemini agent achieved an accuracy of 94.5%, a score that would have placed it at rank #3 among the 434 human participants in the official competition [1]. By contrast, the base Gemini 2.5 Pro model scored only 57.7%, corresponding to rank #92.

This independent evaluation provides strong external validation for the effectiveness and generalizability of our method. First, the success on the IMC demonstrates that our pipeline is not tailored to IMO-style problems but is robust enough to handle the more knowledge-intensive domain of undergraduate mathematics. Second, because the IMC 2025 took place after the public release of our code on GitHub, the competition serves as a pristine testbed, mitigating the concern of data contamination. Finally, the contrast between the agent's performance (rank #3) and that of the base model alone (rank #92) highlights how our verification-and-refinement pipeline translates the latent capabilities of a powerful base model into reliable, high-quality, and competitive mathematical reasoning.

4 Outlook

A direct avenue for enhancing our pipeline's capabilities involves leveraging the more powerful, albeit computationally intensive, variants of the base models used in this study. These include Gemini 2.5 Pro Deep Think, Grok-4 Heavy, and GPT-5 Pro. Integrating them into our pipeline will be a natural and important next step in pushing the frontiers of automated mathematical reasoning.

While our pipeline is model-agnostic, its current implementation operates within a single-model paradigm, where one base LLM serves as both the solver and the verifier. A natural extension of this work is to develop a multi-model collaborative framework that leverages the strengths of different leading LLMs (Gemini 2.5 Pro, Grok-4, and GPT-5). In such a system, each step of our pipeline, from initial solution generation to iterative refinement and verification, would involve two sub-steps: first, each model would work independently to generate a solution or verification report; second, the models would engage in a collective review [14], comparing and critiquing all individual outputs to synthesize a single, consolidated output for that step. This collaborative approach is expected to yield significant benefits. For creative tasks like solution generation, it would pool the diverse reasoning pathways of different models, fostering a richer set of novel ideas. For verification, a subtle error or logic gap missed by one model may be caught by another. By combining the complementary strengths of different models, we believe that such a collaborative system would possess significantly stronger and more reliable mathematical reasoning capabilities.

Acknowledgments and Disclosure of Funding

We thank xAI for providing computational credits for using Grok-4. LY is also supported in part by NSF Grant 2221871 and an Amazon Faculty Award.

References

[1] IMC2025 preliminary results. https://www.imc-math.org.uk/?act=results&by=sum&year=2025.

[2] International Mathematical Olympiad. https://www.imo-official.org.

[3] Not Even Bronze: Evaluating LLMs on 2025 International Math Olympiad. https://matharena.ai/imo/, https://matharena.ai/?comp=imo--imo_2025.

[4] Organisation of the International Mathematical Olympiad. https://imof.co/about-imo/activities/.

[5] With flying colors: Language models ace the International Mathematics Competition. https://matharena.ai/imc/.

[6] R. Agarwal and P. Gaule. Invisible geniuses: Could the knowledge frontier advance faster? *American Economic Review: Insights*, 2(4):409–24, 2020.

[7] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin. Large language models for mathematical reasoning: Progresses and challenges. In N. Falk, S. Papi, and M. Zhang, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics.

[8] AlphaProof and AlphaGeometry teams. AI achieves silver-medal standard solving International Mathematical Olympiad problems. https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/, 2024.

[9] M. Balunović, J. Dekoninck, I. Petrov, N. Jovanović, and M. Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. arXiv:2505.23281.

[10] ByteDance Seed AI4Math. Seed-prover: Deep and broad reasoning for automated theorem proving. arXiv:2507.23726.

[11] Y. Cheng, Y. Chang, and Y. Wu. A survey on data contamination for large language models. arXiv:2502.14425.

[12] Y. Chervonyi, T. H. Trinh, M. Olšák, X. Yang, H. Nguyen, M. Menegali, J. Jung, V. Verma, Q. V. Le, and T. Luong. Gold-medalist performance in solving Olympiad geometry with AlphaGeometry2. arXiv:2502.03544.

[13] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. arXiv:2110.14168.

[14] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] E. First, M. N. Rabe, T. Ringer, and Y. Brun. Baldur: Whole-proof generation and repair with large language models. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2023, page 1229–1241, New York, NY, USA, 2023. Association for Computing Machinery.

[16] Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261.

[17] C. He, R. Luo, Y. Bai, S. Hu, Z. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[18] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[19] R. Kamoi, Y. Zhang, N. Zhang, J. Han, and R. Zhang. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

[20] D. Kim. Looking back on the problem selection committee. https://web.stanford.edu/~dkim04/blog/imo-psc-2023/, 2023.

[21] G. Kim, P. Baldi, and S. McAleer. Language models can solve computer tasks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39648–39677. Curran Associates, Inc., 2023.

[22] T. Luong, E. Lockhart, et al. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-i 2025.

[23] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc., 2023.

[24] H. Mahdavi, A. Hashemi, M. Daliri, P. Mohammadipour, A. Farhadi, S. Malek, Y. Yazdanifard, A. Khasahmadi, and V. G. Honavar. Brains vs. bytes: Evaluating LLM proficiency in olympiad mathematics. In *Second Conference on Language Modeling*, 2025.

[25] D. Paul, M. Ismayilzada, M. Peyrard, B. Borges, A. Bosselut, R. West, and B. Faltings. REFINER: Reasoning feedback on intermediate representations. In Y. Graham and M. Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126, St. Julian's, Malta, Mar. 2024. Association for Computational Linguistics.

[26] I. Petrov, J. Dekoninck, M. D. Lyuben Baltadzhiev, K. Minchev, M. Balunović, N. Jovanović, and M. Vechev. Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad. arXiv:2503.21934.

[27] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[28] P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[29] P.-Y. Wang, T.-S. Liu, C. Wang, Y.-D. Wang, S. Yan, C.-X. Jia, X.-H. Liu, X.-W. Chen, J.-C. Xu, Z. Li, and Y. Yu. A survey on large language models for mathematical reasoning. arXiv:2506.08446.

[30] A. Wei, S. Hsu, and N. Brown. https://x.com/alexwei_/status/1946477742855532918.

[31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[32] Y. Weng, M. Zhu, F. Xia, B. Li, S. He, S. Liu, B. Sun, K. Liu, and J. Zhao. Large language models are better reasoners with self-verification. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore, Dec. 2023. Association for Computational Linguistics.