RIEMANN-BENCH: A BENCHMARK FOR MOONSHOT MATHEMATICS

Surge AI Research

ABSTRACT

Recent AI systems have achieved gold-medal-level performance on the International Mathematical Olympiad, demonstrating remarkable proficiency at competition-style problem solving. However, competition mathematics represents only a narrow slice of mathematical reasoning: problems are drawn from limited domains, require minimal advanced machinery, and can often reward insightful tricks over deep theoretical knowledge. We introduce RIEMANN-BENCH, a private benchmark of 25 expert-curated problems designed to evaluate AI systems on research-level mathematics that goes far beyond the olympiad frontier. Problems are authored by Ivy League mathematics professors, graduate students, and PhD-holding IMO medalists, and routinely took their authors weeks to solve independently. Each problem undergoes double-blind verification by two independent domain experts who must solve the problem from scratch, and yields a unique, closed-form solution assessed by programmatic verifiers. We evaluate frontier models as unconstrained research agents, with full access to coding tools, search, and open-ended reasoning, using an unbiased statistical estimator computed over 100 independent runs per problem. Our results reveal that all frontier models currently score below 10%, exposing a substantial gap between olympiad-level problem solving and genuine research-level mathematical reasoning. By keeping the benchmark fully private, we ensure that measured performance reflects authentic mathematical capability rather than memorization of training data.

1 INTRODUCTION

Five years ago, Surge helped create GSM8K (Cobbe et al., 2021), the first mathematical reasoning benchmark for large language models. At the time, the tasks focused on grade-school math and GSM8K became one of the most widely cited benchmarks because it exposed a fundamental gap between fluent language and actual reasoning.

The frontier has since moved dramatically. The year 2025 marked an important moment for AI and mathematics. Google DeepMind's Gemini with Deep Think scored 35 out of 42 points on the 2025 International Mathematical Olympiad (IMO), officially achieving gold-medal standard as certified by IMO coordinators (DeepMind, 2025). DeepSeekMath-V2 achieved gold-level performance on IMO 2025 and scored 118/120 on Putnam 2024 (DeepSeekMath-V2, 2025). These achievements followed a rapid progression: AlphaProof solved the hardest problem at IMO 2024 (AlphaProof, 2025), and performance on the American Invitational Mathematics Examination (AIME) approached near-perfect accuracy, with o4-mini achieving 99.5% on AIME 2025 using tool access (OpenAI, 2025). The emergence of reasoning-focused models such as OpenAI's o1 (OpenAI, 2024) and DeepSeek-R1 (DeepSeek-AI, 2025), which apply reinforcement learning to develop extended chains of reasoning, has been a key driver of these gains.

The IMO is deliberately limited to four domains: Algebra, Combinatorics, Geometry, and Number Theory. These foundational areas were specifically chosen because they require minimal advanced machinery; calculus, for instance, is strictly excluded. Because the available tools are limited, IMO problems are designed to reward lateral thinking, often hinging on a single key insight that renders the solution tractable. While IMO problems are incredibly clever, they are fundamentally designed to be solved in a few hours using known tools. The distance between this style of problem solving and the sustained, multi-step theoretical reasoning characteristic of professional mathematical research is vast.

This gap motivates a new evaluation paradigm. While benchmarks such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021a), and Omni-MATH (Gao et al., 2024) have progressively raised the difficulty bar, they remain largely confined to competition-style problems. The saturation of existing benchmarks, combined with growing evidence of data contamination in public evaluations (Mirzadeh et al., 2025; Srivastava et al., 2024; Oren et al., 2024), underscores the need for private, rigorously constructed benchmarks at the research frontier.

We introduce RIEMANN-BENCH, a benchmark of 25 extreme-tier mathematical problems designed to evaluate AI not on competition puzzles, but on PhD-level research mathematics. We collaborated with Ivy League mathematics professors, graduate students, and PhD-holding IMO medalists to gather problems they encountered in their own research. These problems routinely took their authors weeks to solve independently. Authors noted that their own graduate students and colleagues would struggle to solve these problems on their own.

Our contributions are:

1. **Research-level mathematical benchmark.** We introduce RIEMANN-BENCH, comprising 25 problems spanning multiple mathematical domains including areas that require understanding of variational principles, measure theory, stability analysis, manifolds, and advanced algebraic structures.
2. **Double-blind, from-scratch verification.** Every problem is independently verified by two domain experts who must solve the problem from scratch before confirming validity.
3. **Rigorous, unconstrained evaluations.** Unlike existing benchmarks that can force models into rigid evaluation loops, RIEMANN-BENCH evaluates true, unconstrained AI research agents with full access to coding tools, search, and open-ended reasoning. We run each frontier model 100 times per problem and compute pass rates using the unbiased estimator of Chen et al. (2021). All models currently score below 10%.
4. **Fully private and uncontaminated.** The dataset is kept strictly private to ensure a fully unbiased evaluation for all frontier labs.

2 RELATED WORK

2.1 MATHEMATICAL REASONING BENCHMARKS

The landscape of mathematical reasoning benchmarks has evolved rapidly, tracing a clear difficulty progression from elementary arithmetic to research-level problems.

Elementary and competition-level benchmarks. GSM8K (Cobbe et al., 2021) introduced 8,500 grade-school math word problems requiring 2–8 reasoning steps, alongside the verifier-based evaluation paradigm. The MATH dataset (Hendrycks et al., 2021a) raised the bar significantly with 12,500 competition-level problems across seven categories sourced from AMC, AIME, and other competitions. When introduced, the best models achieved roughly 7% accuracy; frontier models now exceed 90%, rendering the benchmark effectively saturated. MMLU (Hendrycks et al., 2021b) includes mathematics-related subjects among its 57 domains, spanning elementary through college-level abstract algebra.

Olympiad-level benchmarks. Several recent benchmarks target olympiad-level difficulty. Omni-MATH (Gao et al., 2024) contains 4,428 problems across 33 sub-domains sourced from competitions including USAMO, APMO, and Putnam. OlympiadBench (He et al., 2024) provides 8,476 bilingual problems in mathematics and physics drawn from international olympiads. OlymMATH (OlymMATH, 2025) targets olympiad-level reasoning across multiple difficulty tiers. JEEBench (Arora et al., 2023) uses 515 problems from India's JEE-Advanced examination. MathOdyssey (Fang et al., 2024) contributes 387 expert-crafted problems spanning high school to university level. MathArena (Balunović et al., 2025) evaluates models on recently released competition problems with rigorous contamination controls. The AI Mathematical Olympiad (AIMO) Prize (AIMO Prize, 2023) has further catalyzed progress by awarding prizes for publicly shared models that solve olympiad-level problems.

Graduate and research-level benchmarks. GPQA (Rein et al., 2024) provides 448 expert-crafted multiple-choice questions in physics, chemistry, and biology at a graduate level where domain experts achieve only 65% accuracy. Humanity's Last Exam (Phan et al., 2026) crowdsourced 3,000 expert-level questions across dozens of academic disciplines; frontier models scored below 10% at launch. GHOSTS (Frieder et al., 2023) was among the first benchmarks curated by working mathematicians to target graduate-level mathematics. TheoremQA (Chen et al., 2023) tests knowledge of over 350 theorems across mathematics, physics, and finance. ARB (Sawada et al., 2023) targets graduate and expert-level reasoning across multiple domains. SciBench (Wang et al., 2024) evaluates college-level scientific problem solving with free-response questions drawn from standard textbooks. MathBench (Liu et al., 2024) spans 3,709 problems across five progressive stages from arithmetic to college mathematics.

Formal mathematics benchmarks. MiniF2F (Zheng et al., 2022) contains 488 problems formalized across Lean, Metamath, Isabelle, and HOL Light. ProofNet (Azerbayev et al., 2023) provides 371 parallel examples of formal and natural-language theorem statements from undergraduate textbooks. PutnamBench (Tsoukalas et al., 2024) offers 1,692 hand-constructed formalizations of 640 Putnam competition theorems. DeepSeek-Prover-V2 (DeepSeek-AI, 2025b) recently advanced formal theorem proving by combining reinforcement learning with subgoal decomposition in Lean 4.

FrontierMath. FrontierMath (Glazer et al., 2024), developed by Epoch AI, contains approximately 350 problems organized into four difficulty tiers. Frontier models score close to 40% even on their most challenging tier (Tier 4). RIEMANN-BENCH complements existing benchmarks in several ways: it is independently constructed, fully private, uses double-blind from-scratch expert verification, evaluating models as unconstrained research agents.

2.2 AI PERFORMANCE ON MATHEMATICAL OLYMPIADS

AlphaGeometry (Trinh et al., 2024) solved 25 of 30 historical IMO geometry problems, matching the average gold medalist. AlphaProof (AlphaProof, 2025), combined with AlphaGeometry 2 (Chervonyi et al., 2025), scored 28/42 at IMO 2024 (silver medal; the gold cutoff was 29). By IMO 2025, Gemini with Deep Think became the first AI system officially certified at gold-medal standard by IMO coordinators (DeepMind, 2025). DeepSeekMath-V2 achieved gold-level scores on IMO 2025 and CMO 2024, and scored 118/120 on Putnam 2024 (DeepSeekMath-V2, 2025). In formal mathematics, Axiom Math's AxiomProver solved all 12 problems on the 2025 William Lowell Putnam Mathematical Competition with machine-verified proofs in Lean 4 (Axiom Math, 2025).

In parallel, language model reasoning capabilities have advanced rapidly. Chain-of-thought prompting (Wei et al., 2022) demonstrated that eliciting intermediate reasoning steps substantially improves mathematical performance. Subsequent work on tree-structured reasoning (Yao et al., 2023) and tool-augmented approaches such as PAL (Gao et al., 2023) and ToRA (Gou et al., 2024) further expanded the problem-solving capabilities of language models. The introduction of dedicated reasoning models, beginning with OpenAI o1 (OpenAI, 2024) and followed by DeepSeek-R1 (DeepSeek-AI, 2025), marked a paradigm shift: these systems allocate substantial test-time compute to extended chains of reasoning, yielding dramatic gains on competition mathematics. Domain-specific mathematical training has also proven effective, with Minerva (Lewkowycz et al., 2022) demonstrating early gains through pretraining on mathematical corpora, and more recent systems such as DeepSeek-Math (DeepSeek-AI, 2024), Llemma (Azerbayev et al., 2024), InternLM-Math (Ying et al., 2024), and Qwen2.5-Math (Yang et al., 2024) advancing the state of the art for open-weight mathematical models.

3 BENCHMARK DESIGN

3.1 DESIGN PHILOSOPHY

RIEMANN-BENCH operates in a fundamentally different regime from competition-style benchmarks. While IMO problems are incredibly clever, they are designed to be solved in a few hours using known tools and mathematical machinery. RIEMANN-BENCH operates in the universe of PhD-level research: problems that demand deep domain knowledge, complex theory, and the synthesis of advanced mathematical machinery over long reasoning chains.

The problems are not open-ended conjectures; they have known, verifiable answers. But arriving at those answers demands the kind of multi-step reasoning and mastery of specialized mathematical frameworks that characterizes work at the research frontier. To succeed on RIEMANN-BENCH, an AI system cannot rely on pattern recognition or lateral thinking alone. It must navigate complex abstract definitions, apply advanced theorems, and sustain coherent reasoning across extended chains of increasing complexity.

3.2 PROBLEM CONSTRUCTION

RIEMANN-BENCH comprises 25 problems authored by advanced mathematicians actively engaged in mathematical research. Contributors were asked to draw on problems they encountered in their own research: problems that routinely took them weeks to solve independently. Multiple authors noted that their own graduate students and colleagues would struggle to solve these problems on their own.

Each problem satisfies the following requirements:

- **Unambiguous answer.** Every problem yields a unique, closed-form solution. There is no partial credit and no subjective judgment: the answer is either correct or incorrect.
- **Programmatic verification.** When the solution admits multiple equivalent representations (for example, a rational number that may be expressed in different forms), programmatic verifiers assess correctness automatically.
- **Research-level difficulty.** Problems require deep domain knowledge and multi-step theoretical reasoning that goes substantially beyond what is testable in competition settings.

3.3 VERIFICATION PROTOCOL

Every problem in RIEMANN-BENCH was subjected to a strict double-blind, from-scratch verification protocol:

1. Two independent domain experts, who were not shown the author's solution in advance, are assigned to verify each problem.
2. Each verifier must solve the problem from scratch and arrive at the correct answer through their own reasoning before confirming a problem's validity.
3. The verifiers also assess problem quality, checking for ambiguity, underspecification, and appropriate difficulty calibration.

This double-blind protocol goes substantially beyond the standard practice of relying on problem authors to self-certify their solutions. It provides a strong guarantee that each problem has a unique correct answer that can be independently derived by multiple experts. Problems that failed verification—due to ambiguity, errors, or insufficient difficulty—were revised or excluded.

3.4 PRIVACY

To ensure a fully unbiased evaluation for all frontier labs, the dataset is kept strictly private. Public benchmarks, however well-intentioned, are vulnerable to leakage and contamination (Xu et al., 2024; Zhou et al., 2023). A benchmark that has been seen, even indirectly, is a benchmark that has been compromised. Labs wishing to evaluate their models may submit them through a controlled evaluation service.

3.5 UNCONSTRAINED AGENT EVALUATION

Existing benchmarks can force models into rigid, automated evaluation loops. RIEMANN-BENCH evaluates unconstrained AI research agents. Models are given full access to coding tools (Python interpreter), search capabilities, and open-ended reasoning with no artificial constraints on interaction format or token budget. This design reflects our belief that measuring research-level mathematical capability requires allowing models to operate as they would in a genuine research setting.

4 ILLUSTRATIVE PROBLEM

To convey the character and difficulty of RIEMANN-BENCH, we present one sample problem. This problem illustrates several key properties of the benchmark: it involves advanced mathematical objects, requires deep familiarity with specialized theory, and demands sustained multi-step reasoning that a domain expert estimated would take 40–50 hours to complete from scratch.

Problem overview. The problem concerns the classification of multibasic A-modules over the ring of Hahn series with real-valued valuation and residue field $\mathbb{F}_2$. The field K of Hahn series in indeterminate t with value group $\mathbb{R}$ is considered as a module over its subring A of elements with non-negative valuation. Special A-modules, termed *basic* (quotients of submodules of K) and *multibasic* (finite direct sums of basic modules), are defined, with the property that every multibasic A-module has a unique decomposition into a direct sum of basic submodules. The problem asks for the number of distinct isomorphism classes of multibasic A-modules M satisfying three structural conditions involving the endomorphism ring and a dimension function on associated $\mathbb{F}_2$-vector spaces.

Discussion. Hahn series with real-valued valuation are formal infinite series in which the exponents may be any real numbers. A key insight in the solution is that submodules of the Hahn series field behave very simply with respect to the valuation: any submodule is determined by which powers t^q it contains, forcing the submodule to correspond to a cut in $\mathbb{R}$. As a result, every basic module L/N must come from a small list of canonical possibilities. Because multibasic modules decompose uniquely into basic pieces, the classification problem reduces to determining which combinations of these building blocks are allowed.

The three conditions in the problem then act as filters, each eliminating a different family of candidates through a qualitatively distinct algebraic mechanism. The solution draws on a diversity of mathematical ideas: classifying A-submodules of K via the valuation, applying the tensor-hom adjunction to determine how tensor products and hom functors interact with multibasic modules, using ring-theoretic properties to constrain which basic summands can appear, and finally reducing to a finite case analysis with a combinatorial count.

5 EXPERIMENTAL SETUP

5.1 MODELS EVALUATED

We evaluated major frontier AI models. All models were evaluated through their respective APIs, with full access to coding tools (Python interpreter), search capabilities, and open-ended reasoning. No artificial constraints were imposed on interaction format or token budget. This setup ensures that measured performance gaps reflect genuine mathematical reasoning limitations rather than implementation bottlenecks.

5.2 EVALUATION PROTOCOL

For each problem, we ran every model 100 times independently and computed pass rates using the unbiased statistical estimator introduced by Chen et al. (2021). Given n total samples and c correct samples, the pass@k estimator is:

$$\text{pass@}k = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}$$

This estimator provides unbiased measurements of model capability at any sampling budget k, computed from a fixed pool of $n = 100$ independent attempts. The resulting difficulty assessments are not based on a small number of attempts or selected runs; they reflect stable, reproducible measurements.

6 RESULTS

6.1 OVERALL PERFORMANCE

Table 1 and Figure 1 present our primary results across all 25 problems.

Table 1: Frontier model performance on Riemann-Bench. Pass rates are computed from 100 independent runs per problem using the unbiased estimator of Chen et al. (2021). All models were evaluated as unconstrained agents with access to coding tools and search.

Model	**Lab**	**pass@1 (%)**
Gemini 3.1 Pro	Google	6
Claude Opus 4.6	Anthropic	6
Gemini 3 Pro	Google	4
Kimi K2.5	Moonshot AI	4
DeepSeek V3.2	DeepSeek	3
GPT 5.2	OpenAI	2
Claude Opus 4.5	Anthropic	2

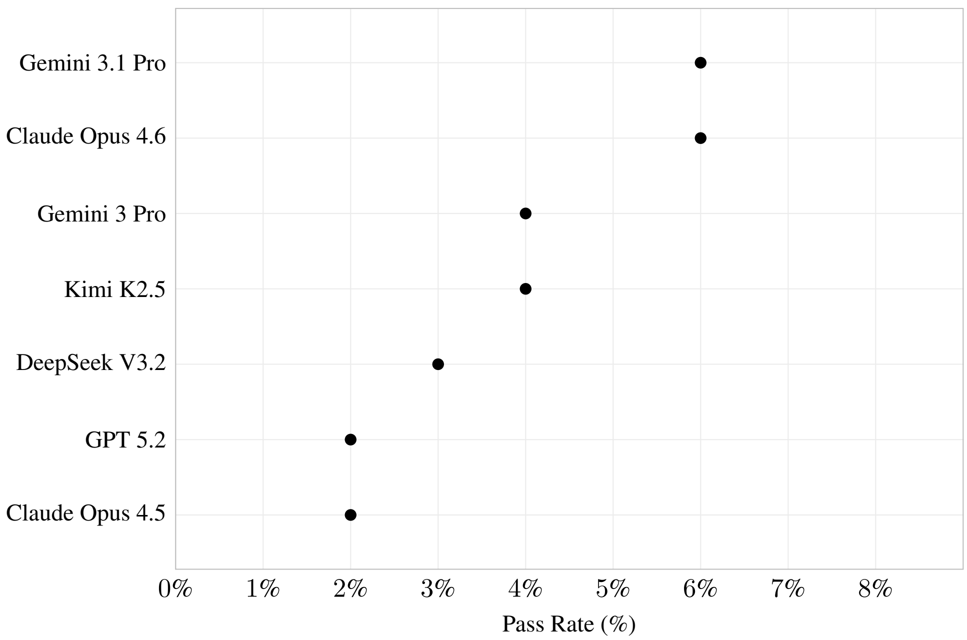

Figure 1: pass@1 across frontier models on Riemann-Bench. All models score below 10%, confirming that research-level mathematics remains substantially beyond current capabilities.

The central finding is stark: **all frontier models currently score below 10% on Riemann-Bench, even when operating as unconstrained research agents with full access to coding tools and search.** This is in sharp contrast to olympiad-level benchmarks, where the same generation of models approaches or exceeds human gold-medal performance. The gap confirms that research-level mathematics, the kind of sustained, multi-step theoretical reasoning that characterizes PhD-level work, remains beyond current model capabilities.

6.2 Comparison with Competition-Level Performance

To contextualize these results, we note the performance of the same model generation on competition-level mathematics. The models evaluated here, or their close variants, achieve near-perfect scores on AIME problems and gold-medal-level performance on IMO problems. The dramatic drop from near-100% on AIME to below 10% on Riemann-Bench illustrates the qualitative difference between competition mathematics and research-level problems. Competition problems, however difficult, can often be resolved through a single key insight applied with elementary tools. Riemann-Bench problems require sustained theoretical reasoning over weeks of effort, drawing on specialized knowledge that extends well beyond the competition canon.

6.3 Analysis of a Representative Failure Mode

Beyond aggregate pass rates, qualitative analysis of model failures reveals important patterns in how current systems break down on research-level mathematics. We present a representative failure on the illustrative problem from Section 4 to demonstrate an important class of errors.

Model failure. Rather than working within the A-module framework specified by the problem, the model reinterpreted the entire problem in terms of an inapplicable theory of "generalized scales." Specifically, it incorrectly treated conditions (i) and (ii)—constraints on the endomorphism ring of M, as the definition of a "basic scale," and misinterpreted condition (iii) as specifying the "support" of M, when in fact the relevant notion of support is intrinsic to the Hahn series construction. To justify its reasoning, the model fabricated a nonexistent classification theorem, attributing it to a fictitious reference ("M. Getz, Theorem 4.14 on Generalized Scales"). Applying this fabricated theorem, the model arrived at an answer of 2^{299}, which is off by orders of magnitude from the correct answer.

Broader pattern. This failure exemplifies a recurring pattern observed across RIEMANN-BENCH evaluations: when confronted with problems requiring specialized theoretical frameworks, models may substitute a superficially related but inapplicable framework and fabricate supporting results to complete the reasoning chain. The model's output reads as structurally coherent: it identifies the problem as a classification task, proposes a theoretical framework, invokes a theorem, and computes a numerical answer, but the entire reasoning chain is built on a misidentified foundation.

7 Discussion

7.1 Why Competition Math Is Not Enough

The contrast between olympiad-level and research-level performance reveals a fundamental distinction in mathematical reasoning. Problems in RIEMANN-BENCH demand deep familiarity with specialized theory, the ability to chain together multiple advanced results, and sustained computation through complex algebraic or analytic manipulations. A model that can identify the key insight in an IMO problem may nonetheless be unable to navigate the Eynard–Orantin topological recursion or classify multibasic modules over Hahn series rings.

This distinction carries important implications for how we interpret benchmark saturation. The saturation of competition-level benchmarks does not imply that mathematical reasoning is solved. It indicates that a particular style of mathematical reasoning, one that rewards lateral thinking with more elementary tools, is within reach of current systems. The broader and deeper landscape of mathematics, encompassing the specialized theory and extended reasoning chains required for research-level work, remains largely beyond current capabilities.

7.2 Implications for AI-Assisted Mathematical Research

The results carry both encouraging and cautionary implications for the prospect of AI systems contributing to mathematical research. The rapid generational improvement observed on competition mathematics suggests that continued scaling and targeted training can yield meaningful progress. On the other hand, performance below 10% on RIEMANN-BENCH means that even the best models fail on the vast majority of research-level problems, making current systems unreliable as autonomous mathematical reasoning agents.

A more realistic near-term application may be AI-assisted research, in which human mathematicians use AI systems as computational assistants for specific subtasks while verifying outputs against their own expertise. This mirrors the trajectory observed in other domains of AI deployment, where practitioners deliberately constrain agent autonomy to maintain reliability (Pan et al., 2025). Tools such as Lean Copilot (Song et al., 2024) exemplify this collaborative approach in the context of formal theorem proving.

7.3 Toward Moonshot Mathematics

RIEMANN-BENCH problems, however difficult, still have known solutions. The true moonshots of mathematical research require formulating conjectures, building novel frameworks, and navigating spaces in which the existence of an answer is itself unknown. We view RIEMANN-BENCH as a necessary intermediate evaluation along this trajectory. Reliable performance on research-level problems with known solutions is a prerequisite for any system aspiring to contribute to open mathematical research.

8 CONCLUSION

We introduced RIEMANN-BENCH, a private benchmark of 25 extreme-tier mathematical problems for research-level reasoning. Our principal findings are:

- All frontier models currently score below 10% on RIEMANN-BENCH, revealing a vast gap between olympiad-level problem solving and research-level mathematical reasoning.
- The double-blind, from-scratch expert verification protocol and fully private evaluation design ensure that measured performance reflects genuine mathematical capability rather than memorization.
- Evaluating models as unconstrained research agents rather than constraining them to rigid evaluation loops, provides a more faithful measure of AI's capacity for open-ended mathematical reasoning.
- Qualitative analysis of failures reveals that models can substitute inapplicable theoretical frameworks and fabricate supporting results, producing structurally coherent but substantively wrong reasoning chains.

Having built the baseline the field relies on with GSM8K, we are now defining its ceiling with RIEMANN-BENCH. AI's success on the IMO marks a beginning, not an end. RIEMANN-BENCH provides a rigorous, contamination-resistant measurement of progress toward the mathematical moonshots that matter.

ACKNOWLEDGMENTS

We thank the mathematicians and researchers who contributed problems to RIEMANN-BENCH and the domain experts who participated in the double-blind verification protocol. Their expertise and rigor are the foundation of this benchmark.

REFERENCES

Hubert, T., Baudisin, A., Banerjee, A., et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 2025.

Arora, D., Singh, H. M., and Mausam. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. In *Proceedings of EMNLP*, 2023.

Axiom Math. AxiomProver Solves All Problems at Putnam 2025. https://axiommath.ai/territory/from-seeing-why-to-checking-everything, December 2025.

Azerbayev, Z., Piotrowski, B., Schoelkopf, H., et al. ProofNet: Autoformalizing and Formally Proving Undergraduate-Level Mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Azerbayev, Z., Schoelkopf, H., Paster, K., et al. Llemma: An Open Language Model For Mathematics. In *Proceedings of ICLR*, 2024.

Balunović, M., Beutel, L., Sairam, P., Vechev, M., and Giannakopoulos, N. MathArena: Evaluating LLMs on Uncontaminated Math Competitions. In *Advances in NeurIPS Datasets and Benchmarks*, 2025.

Chen, M., Tworek, J., Jun, H., et al. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021.

Chen, W., Yin, M., Ku, M., et al. TheoremQA: A Theorem-driven Question Answering Dataset. In *Proceedings of EMNLP*, 2023.

Chervonyi, Y., et al. Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2. *arXiv preprint arXiv:2502.03544*, 2025.

Cobbe, K., Kosaraju, V., Bavarian, M., et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Google DeepMind. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. Blog post, July 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Nature*, 645:633–638, 2025.

Shao, Z., Wang, P., Zhu, Q., et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Ren, X., et al. DeepSeek-Prover-V2: Advancing Formal Mathematical Reasoning via Reinforcement Learning for Subgoal Decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

Shao, Z., Luo, Y., Lu, C., et al. DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning. *arXiv preprint arXiv:2511.22570*, 2025.

Fang, F., Mei, X., Miao, Z., et al. MathOdyssey: Benchmarking Mathematical Problem-Solving Skills in Large Language Models Using Odyssey Math Data. *arXiv preprint arXiv:2406.18321*, 2024.

Frieder, S., Pinchetti, L., Griffiths, R.-R., et al. Mathematical Capabilities of ChatGPT. In *Advances in NeurIPS*, 2023.

Gao, L., Madaan, A., Zhou, S., et al. PAL: Program-aided Language Models. In *Proceedings of ICML*, 2023.

Gao, B., Song, Y., et al. Omni-MATH: A Universal Olympiad Level Mathematic Benchmark For Large Language Models. *arXiv preprint arXiv:2410.07985*, 2024.

Glazer, E., Erdil, E., Besiroglu, T., et al. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Gou, Z., Shao, Z., Gong, Y., et al. ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving. In *Proceedings of ICLR*, 2024.

He, C., Luo, R., Bai, Y., et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of ACL*, 2024.

Hendrycks, D., Burns, C., Kadavath, S., et al. Measuring Mathematical Problem Solving With the MATH Dataset. In *Advances in NeurIPS*, 2021.

Hendrycks, D., Burns, C., Basart, S., et al. Measuring Massive Multitask Language Understanding. In *Proceedings of ICLR*, 2021.

Lewkowycz, A., Andreassen, A., Dohan, D., et al. Solving Quantitative Reasoning Problems with Language Models. In *Advances in NeurIPS*, 2022.

Liu, H., Zheng, Z., et al. MathBench: Evaluating the Theory and Application Proficiency of LLMs with a Hierarchical Mathematics Benchmark. In *Findings of ACL*, 2024.

Mirzadeh, I., Alizadeh-Vahid, K., et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *Proceedings of ICLR*, 2025.

OlymMATH authors. Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models. *arXiv preprint arXiv:2503.21380*, 2025.

OpenAI. OpenAI o1 System Card. *arXiv preprint arXiv:2412.16720*, 2024.

OpenAI. OpenAI o3 and o4-mini System Card. Technical report, April 2025.

Oren, Y., Meister, N., Chatterji, N., Lauj, F., and Hashimoto, T. Proving Test Set Contamination in Black Box Language Models. In *Proceedings of ICLR*, 2024.

Pan, M. Z., Arabzadeh, N., Cogo, R., et al. Measuring Agents in Production. *arXiv preprint arXiv:2512.04123*, 2025.

Phan, L., et al. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026.

AI Mathematical Olympiad Prize. https://aimoprize.com/, 2023.

Rein, D., Hou, B. L., Stickland, A. C., et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *Proceedings of ICLR*, 2024.

Sawada, T., Paleka, D., Havrilla, A., et al. ARB: Advanced Reasoning Benchmark for Large Language Models. *arXiv preprint arXiv:2307.13692*, 2023.

Song, P., Yang, K., and Anandkumar, A. Lean Copilot: LLMs as Copilots for Theorem Proving in Lean. *arXiv preprint arXiv:2404.12534*, 2024.

Srivastava, M., et al. Functional Benchmarks for Robust Evaluation of Reasoning Performance, and the Reasoning Gap. *arXiv preprint arXiv:2402.19450*, 2024.

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

Tsoukalas, G., Jasber, J., et al. PutnamBench: Evaluating Neural Theorem-Provers on the Putnam Mathematical Competition. *arXiv preprint arXiv:2407.11214*, 2024.

Wang, X., Hu, Z., Lu, P., et al. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. In *Proceedings of ICML*, 2024.

Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in NeurIPS*, 2022.

Xu, R., et al. Benchmark Data Contamination of Large Language Models: A Survey. *arXiv preprint arXiv:2406.04244*, 2024.

Yang, A., Zhang, B., Hui, B., et al. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv preprint arXiv:2409.12122*, 2024.

Yao, S., Yu, D., Zhao, J., et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in NeurIPS*, 2023.

Ying, H., Zhang, S., et al. InternLM-Math: Open Math Large Language Models Toward Verifiable Reasoning. *arXiv preprint arXiv:2402.06332*, 2024.

Zheng, K., Han, J. M., and Polu, S. MiniF2F: A Cross-System Benchmark for Formal Olympiad-Level Mathematics. In *Proceedings of ICLR*, 2022.

Zhou, K., Zhu, Y., et al. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.01964*, 2023.