

Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models

Haoxiang Sun¹, Yingqian Min², Zhipeng Chen²,
Wayne Xin Zhao^{2,3}✉, Ji-Rong Wen²

¹ School of Information, Renmin University of China

² Gaoling School of Artificial Intelligence, Renmin University of China

³ Beijing Key Laboratory of Research on Large Models and Intelligent Governance
{hxiang.sun, batmanfly}@gmail.com

Abstract

The rapid advancement of large reasoning models has saturated existing math benchmarks, underscoring the urgent need for more challenging evaluation frameworks. To address this, we introduce **OlymMATH**, a rigorously curated, Olympiad-level math benchmark comprising 350 problems, each with parallel English and Chinese versions. OlymMATH is the first benchmark to unify dual evaluation paradigms within a single suite: (1) *natural language evaluation* through **OlymMATH-EASY** and **OlymMATH-HARD**, comprising 200 computational problems with numerical answers for objective rule-based assessment, and (2) *formal verification* through **OlymMATH-LEAN**, offering 150 problems formalized in Lean 4 for rigorous process-level evaluation. All problems are manually sourced from printed publications to minimize data contamination, verified by experts, and span four core domains. Extensive experiments reveal the benchmark’s significant challenge, and our analysis also uncovers consistent performance gaps between languages and identifies cases where models employ heuristic “guessing” rather than rigorous reasoning. To further support community research, we release 582k+ reasoning trajectories, a visualization tool, and expert solutions at <https://github.com/RUCAIBox/OlymMATH>.

1 Introduction

The advent of large language models (LLMs) (Zhao et al., 2023) has marked a significant leap forward in the capabilities of artificial intelligence, with mathematical reasoning emerging as a pivotal and demanding area of research (DeepSeek-AI et al., 2025a; OpenAI, 2024; Chen et al., 2025). Recently, the evaluation and enhancement of mathematical reasoning abilities have become a central focus in the development of LLMs (Yang et al., 2024).

Effective assessment of LLM reasoning necessitates *reliable* and *verifiable* evaluation benchmarks. Reliability requires accurately designed problems with unambiguous solutions and minimized data contamination risk, ensuring trustworthy evaluation. For verifiability, two paradigms have emerged: (1) *numerical-answer benchmarks* using rule-based verification (e.g., sympy), which offer scalability but cannot assess reasoning quality; and (2) *formal proof benchmarks* using theorem provers (e.g., Lean, Isabelle), which provide rigorous process-level verification but require specialized formalization. An ideal benchmark suite should leverage both paradigms for comprehensive evaluation.

Existing benchmarks in both paradigms leave certain dimensions only partially addressed, and these gaps are not straightforwardly closed by recombining available resources. Among numerical-answer benchmarks, Olympiad-difficulty collections may not yet provide sufficient scale for robust statistical conclusions or enough headroom for the strongest current models (see Table 1); LLM-as-judge evaluation can be susceptible to evaluator hallucination, and reference solutions are occasionally incomplete. Among formal proof benchmarks, available datasets are English-only and drawn from well-known competitions with substantial online presence, warranting attention to potential contamination. More broadly, most existing benchmarks center on English, leaving multilingual reasoning comparatively less explored. Jointly ensuring consistent quality and low contamination across all these dimensions remains an open problem, as assembling web-crawled sources from different origins risks reintroducing the issues that motivate this work.

To bridge the gap, we present **OlymMATH**, a rigorously curated, bilingual (English and Chinese) benchmark for Olympiad-level reasoning, comprising 350 unique problems organized into three non-overlapping subsets: **OlymMATH-EASY** and

✉Corresponding author.

OlymMATH-EASY / HARD

Problem-EN: Find the remainder of $\sum_{k=0}^{1234} \binom{2016 \times 1234}{2016k}$ modulo 2017^2 (provide the value in the range $[0, 2017^2)$).
Answer: 1581330. **Subject:** Number Theory.

OlymMATH-LEAN

Subject: Number Theory. **Formal Statement:**
theorem to_prove
(n : Nat) (p : Nat) (hp : p.Prime)
(hdiv : p | 2^n + 1) : p % 8 != 7 := by sorry

Figure 1: Examples from our OlymMATH dataset.

OlymMATH-HARD contain 100 computational problems each, split into *easy* and *hard* levels with parallel bilingual versions, requiring precise numerical answers for reliable and rule-based sympy verification. Additionally, **OlymMATH-LEAN** provides a separate set of 150 problems formalized in Lean 4, accompanied by bilingual natural language statements and solutions, enabling rigorous evaluation of automated theorem proving capabilities. Unlike proof-based benchmarks that rely on unreliable LLM-as-a-judge evaluation, OlymMATH-LEAN leverages the Lean language for fully automated and mathematically rigorous formal verification. Meanwhile, to prevent data leakage, problems were manually sourced from printed publications and verified by experts. The benchmark covers four major mathematical fields and adheres to the MATH or miniF2F dataset format for compatibility (see Figure 1).

By leveraging OlymMATH, we conduct extensive experiments to evaluate the performance of state-of-the-art models. The results underscore our benchmark’s difficulty, with advanced models like DeepSeek-R1, o3-mini, and Gemini 2.5 Pro Exp achieving only 19.5%, 31.2%, and 58.4% accuracy, respectively, on OlymMATH-HARD (EN), indicating Olympiad-level math remains a significant challenge necessitating further research. Our bilingual comparison showed a consistent performance gap, with higher accuracy on English problems versus Chinese, highlighting the need for multilingual evaluation. Furthermore, case studies revealed models sometimes use heuristic “guessing” to reach answers without rigorous proofs. This underscores the importance of process-level inspection for accurate LLM capability assessment.

In summary, our contributions are as follows.

- We introduce **OlymMATH**, the first Olympiad-level mathematical benchmark that unifies natural

language problems and formal theorem proving within a single bilingual suite. OlymMATH comprises 350 unique problems, each available in both English and Chinese: **OlymMATH-EASY** and **OlymMATH-HARD** provide 200 computational problems with sympy-verifiable numerical answers, while **OlymMATH-LEAN** offers 150 problems formalized in Lean 4 for process-level verification—bridging the gap between outcome-based and reasoning-based evaluation.

- Extensive experiments validate OlymMATH’s reliability and strong discriminative power, while revealing critical model limitations including EN-ZH performance gaps and heuristic “guessing” that bypasses rigorous reasoning.

- We open-source 582,400 reasoning trajectories from 28 models, a visualization tool, and expert solutions to facilitate community research.

2 Related Work

Natural Language Math Benchmarks. The first paradigm of math reasoning evaluation relies on numerical-answer benchmarks with rule-based verification due to their simplicity and scalability. Early benchmarks such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) have been pivotal in advancing LLM reasoning capabilities (Fang et al., 2024; Arora et al., 2023), but are now largely saturated by slow-thinking models enhanced by long chain-of-thought fine-tuning (Min et al., 2024) or reinforcement learning scaling (DeepSeek-AI et al., 2025a).

More challenging benchmarks face different limitations. The AIME dataset offers increased difficulty but suffers from small scale (e.g., 30 problems from AIME 2025), compromising statistical reliability—a single problem shift can change accuracy by 3.33% (Hochlehnert et al., 2025), with binomial standard errors approximately $2.6\times$ larger than those from a 200-problem benchmark. Moreover, rapidly improving models are approaching its measurement ceiling (e.g., Gemini 2.5 Pro achieving 92% Pass@1 on AIME 2024), and its English-only focus neglects multilingual evaluation. OlympiadBench (He et al., 2024) provides a larger collection of problems, but its overall difficulty remains limited—even a 1.5B model (DeepScaleR-Preview (Luo et al., 2025)) achieves 50.0% accuracy, indicating insufficient challenge for evaluating state-of-the-art reasoning models. Omni-MATH (Gao et al., 2025) increases prob-

lem count via web crawling from AoPS, elevating data leakage risks, while its prevalent proof-based problems require LLM-as-judge evaluation rather than rule-based verification, reducing dependability. PolyMath (Wang et al., 2025b) directly sources from widely publicized competitions (e.g., AIME, CNMO, IMO) and existing datasets (e.g., MGSM (Shi et al., 2023), P-MMEval (Zhang et al., 2025), HLE (Phan et al., 2025)), risking high data leakage. AMO Bench (An et al., 2025) targets advanced Olympiad problems but has limited scale.

In contrast, OlymMATH provides 200 challenging Olympiad-level problems with lower contamination risk through manual curation from printed publications, larger scale for statistical reliability, and bilingual versions for thorough evaluation.

Formal Language Math Benchmarks. Beyond assessing final answers, understanding *how* models arrive at solutions—distinguishing rigorous derivation from heuristic shortcuts—is equally important. Formal theorem proving benchmarks address this by requiring machine-checkable proofs. miniF2F (Zheng et al., 2022) provides Olympiad-level problems formalized in multiple proof assistants, while ProofNet (Azerbayev et al., 2023) focuses on undergraduate mathematics. FIMO (Liu et al., 2023) and PutnamBench (Tsoukalas et al., 2024) target competition mathematics at different levels. However, existing formal benchmarks typically source from well-known competitions (e.g., IMO shortlist for FIMO, Putnam for PutnamBench) with high online exposure, facing similar contamination risks. Moreover, all existing formal benchmarks are English-only. OlymMATH-LEAN addresses these gaps with bilingual Lean 4 formalizations sourced from printed publications.

3 Benchmark Construction

In this section, we describe OlymMATH in detail, including its construction methodology, problem composition, categorical distribution, and evaluation approach. Table 1 presents a comparison with existing mathematical reasoning benchmarks. Existing benchmarks typically focus on either *natural language problems* with numerical answers or *formal theorem proving*, but not both. OlymMATH bridges this gap by being the first Olympiad-level benchmark to integrate both paradigms within a unified bilingual framework: OlymMATH-EASY and OlymMATH-HARD provide 200 natural language problems requiring precise numerical an-

(a) Natural Language Benchmarks			
Name	# Prob. (# Lang.)	Eval.	Difficulty
AIME 24,25	30 (1)	Rule	Olympiad
HMMT	30 (1)	Rule	Olympiad
USAMO 2025	6 (1)	LLM	Olympiad
OlympiadBench	2133 (2)	Rule	Olympiad
Omni-MATH	4428 (1)	LLM	Olympiad
PolyMath	500 (18)	Rule	Olympiad
AMO Bench	50 (1)	Rule & LLM	Olympiad
EASY (Ours)	100 (2)	Rule	Olympiad
HARD (Ours)	100 (2)	Rule	Olympiad

(b) Formal Language Benchmarks			
Name	# Prob.	Lang.	Difficulty
miniF2F	488	EN	Olympiad
ProofNet	371	EN	Undergrad
FIMO	149	EN	Olympiad
PutnamBench	640	EN	Undergrad Comp.
LEAN (Ours)	150	EN & ZH	Olympiad

Table 1: Comparison of existing benchmarks. EN and ZH denote English and Chinese, respectively.

swers for scalable rule-based verification, while OlymMATH-LEAN offers 150 problems with Lean 4 formalizations enabling rigorous process-level verification. This dual-paradigm design allows comprehensive assessment of both outcome correctness and reasoning rigor, addressing the limitations of relying on either paradigm alone.

3.1 Contamination Analysis & Verification

Contamination Analysis Constructing a reliable benchmark requires mitigating data contamination. OlymMATH comprises 350 problems curated from printed resources (specialized magazines and textbooks), intentionally excluding online repositories to minimize prior digital exposure, unlike existing benchmarks drawing from well-known competitions (e.g., FIMO using IMO shortlist, PutnamBench using Putnam, Omni-MATH using AoPS).

For quantitative leakage analysis, we followed Omni-MATH, using n -gram accuracy metric (Xu et al., 2024): for each sample, the problem and answer are concatenated; 5 starting points are uniformly sampled; and the model’s ability to predict the subsequent 5-gram is evaluated. Leakage risk is quantified by comparing n -gram accuracy on the original dataset against 3 LLM-rewritten versions (Gemini 2.5 Flash Preview Thinking (Google, 2025b)), with the normalized difference δ indicat-

Model (Base)	Lang.	PolyMath	OlymMATH
InternLM2-Math-7B (Ying et al., 2024)	EN	34.84%	0.90%
	ZH	12.29%	0.88%
Qwen2.5-7B (Qwen et al., 2025)	EN	38.81%	17.59%
	ZH	10.27%	3.42%

Table 2: Results of leakage analysis. The lower value is bolded. OlymMATH exhibits lower δ values than PolyMath per language, indicating a lower leakage risk.

ing model familiarity with original versus rewritten data. Since δ 's absolute value depends on the rewriting LLM, assessing leakage risk requires *relative comparison* of δ between benchmarks. Results in Table 2 show lower contamination risk for OlymMATH than PolyMath, establishing OlymMATH as a more reliable benchmark for evaluating LLMs' true mathematical capabilities.

Verification To enhance dataset reliability, we invited a China Mathematical Olympiad silver medalist and two provincial first-prize winners to verify and revise the problems and solutions in OlymMATH-EASY and HARD. Since the answers to the problems were already provided, the verification difficulty was reduced, making the expertise of reviewers sufficient for this task. Each problem was reviewed by at least two reviewers. Additionally, we have published official solutions for challenging problems for community oversight.

For OlymMATH-LEAN, we leverage the Lean server for automatic verification. Raw problems and solutions are first cleaned by Claude Opus 4.5 (Anthropic, 2025b) for L^AT_EX format correction, then undergo three independent verification rounds using DeepSeek V3.2 Speciale (DeepSeek-AI et al., 2025b) to check translation accuracy, statement precision, and solution rigor. A Claude Opus 4.5-based agent (see Appendix A.2 for details) then iteratively interacts with a Kimina Lean REPL server (Santos et al., 2025) in an isolated sandbox, refining code based on compiler feedback until successful compilation. Compiled formalizations are validated by three independent Gemini 3.0 Flash (Google, 2025a) calls for semantic alignment, and formalizations containing axiom declarations receive additional human expert reviews.

3.2 Problem Categories and Distribution

OlymMATH problems span four key high-school Olympiad mathematical fields—algebra, geometry, number theory, and combinatorics—classified by

Category	# Problems		
	EASY	HARD	LEAN
Algebra (Alg.) <i>Inequality, Trigonometry, etc.</i>	25	25	79
Geometry (Geo.) <i>Solid & Analytic Geometry, etc.</i>	33	25	15
Number Theory (Num.) <i>Diophantine Equation, etc.</i>	13	25	42
Combinatorics (Com.) <i>Graph Theory, Permutation, etc.</i>	29	25	14
Total	100	100	150

Table 3: The distribution of OlymMATH by category.

human experts for reliability. Problems are selected for their challenge, suitability for simple-answer verification, and topic diversity (e.g., inequalities, sequences, and more in algebra). Figure-based problems within this set are text-reformulated for LLM compatibility, with non-convertible ones excluded (e.g., Figure 5 in Appendix).

For refined evaluation, computational problems are categorized by difficulty: *easy*, designed to challenge standard prompting in mainstream models, and *hard*, tailored to test advanced reasoning (e.g., slow-thinking modes) in state-of-the-art models. Additionally, OlymMATH-LEAN provides 150 problems with Lean 4 (Mathlib v4.24.0) formalizations for process-level verification. The distribution details are described in Table 3.

3.3 Format and Verification Methodology

OlymMATH adopts MATH and miniF2F dataset format (see Figure 1) for seamless integration with existing pipelines and enhancing clarity and processing efficiency. All problems are text-based, including geometry problems reformulated from diagrams to align with LLM evaluation, as mentioned previously. For consistent, objective assessment, answers to computational problems are restricted to real numbers and intervals, e.g., " $[\sqrt{33}, +\infty)$ ", while excluding ambiguous formats such as set operations, variables, complex numbers, and texts (see Table 11 in Appendix for details). This design enables reliable sympy-based and formal Lean server verification.

To make the evaluation more challenging, OlymMATH includes problems with multiple numerical answers. These problems are modified to require a summary of all potential outcomes (e.g., sums, sums of squares; see Figure 6 in Appendix). This method effectively assesses whether models can

consider all possible answers, thereby providing a robust evaluation of their reasoning capabilities.

3.4 Bilingual Extension

Originating from Chinese-language problems, OlymMATH provides both original Chinese and translated English versions for bilingual evaluation. Our translation pipeline employs Claude Sonnet 3.7 (Anthropic, 2025a) for initial translation, iterative refinement with GPT-4o (OpenAI et al., 2024), and human verification by two expert annotators to ensure mathematical accuracy and linguistic fluency. OlymMATH-LEAN similarly provides bilingual natural language statements alongside Lean formalizations, supporting research in multilingual reasoning and informal-formal translation.

4 Experiments

In this section, we assess the performance of leading reasoning models using OlymMATH and provide a detailed analysis of their capabilities.

4.1 Natural Language: EASY & HARD Subset

We first evaluate models on the natural language subsets, where problems require numerical answers verified via rule-based matching.

4.1.1 Experimental Setup

Models. We assess representative LLMs for a thorough evaluation. For open-source models, we investigated recent work on reasoning models, and evaluated DeepSeek-R1 (DeepSeek-AI et al., 2025a), STILL-3-Preview (Team, 2025d), DeepScaleR-Preview (Luo et al., 2025), QwQ (Team, 2025c), Light-R1 (Wen et al., 2025), OpenThinker2 (Team, 2025a), Skywork-OR1 (He et al., 2025), GLM-Z1-Air (GLM et al., 2024), AceMath-RL (Liu et al., 2025), OpenMath-Nemotron (Moshkov et al., 2025), and Qwen3 (Team, 2025b). For closed-source models, we evaluate o3-mini (high) (OpenAI, 2025) and Gemini 2.5 Pro Exp 0325 (Google, 2025).

Evaluation Details. Our evaluation pipeline generates 64 responses per problem for each model, except for resource-intensive models (*i.e.*, OpenMath-Nemotron-32B, Qwen3-235B-A22B, GLM-Z1-Air, DeepSeek-R1, o3-mini (high), and Gemini 2.5 Pro Exp), which are limited to 8 samples due to resource limitations and the relatively large scale of our dataset. For Pass@1, we compute mean accuracy across all sampled responses; for

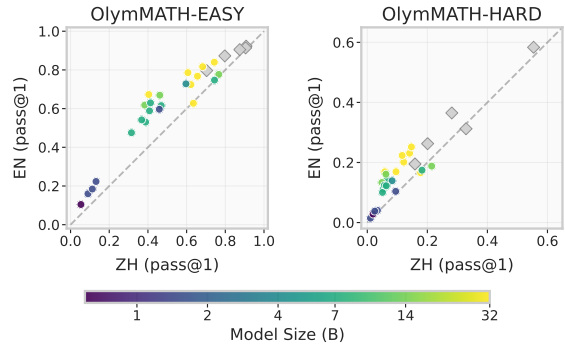


Figure 2: Pass@1 on OlymMATH EN (y) vs. ZH (x), the dashed line shows parity. Points above favor EN, below favor ZH. Solid circles (local dense models, colored by size) indicate larger models trend towards higher accuracy. Diamonds are MoE or closed-source models.

Cons@64 and Cons@8, we apply majority voting to determine a consensus answer per problem. Following established practices (DeepSeek-AI et al., 2025a; Team, 2025c), locally-evaluated models use temperature = 0.6, top_p = 0.95, min_p = 0, and max_tokens = 32768, while API-evaluated models (*i.e.*, GLM-Z1-Air, DeepSeek-R1, o3-mini (high), and Gemini 2.5 Pro Exp) use maximum available max_tokens to fully leverage their reasoning capabilities. We open-source all 582,400 samples, an online visualization tool, and standard solutions for challenging problems to support community analysis of LLM reasoning patterns and characteristics (see Appendix A.1 for further information).

4.1.2 Evaluation Results

In this part, we present the evaluation results of OlymMATH (EN) and OlymMATH (ZH) in Table 4. Due to space constraints, we include only representative models in the main text, and full results are provided in Tables 9 and 10 in Appendix.

First, we observe that all tested models exhibit relatively poor performance, with even OpenAI o3-mini (high) and Gemini 2.5 Pro Exp achieving only 31.2% and 58.4% on OlymMATH-HARD (EN). This underscores the high overall difficulty of our benchmark, which demands stronger reasoning abilities and a deeper understanding of mathematical knowledge. In contrast, the performance of these advanced reasoning models on OlymMATH-EASY (EN) is more modest and comparable to that on AIME 2024, suggesting that OlymMATH-EASY is well-suited for evaluating the capabilities of less advanced reasoning models.

Second, by comparing the performance of LLMs

Model	OlymMATH-HARD										OlymMATH-EASY									
	Alg.		Geo.		Num.		Com.		Avg.		Alg.		Geo.		Num.		Com.		Avg.	
	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k
<i>English (EN)</i>																				
Qwen3 (0.6B, Think)	2.5	0.0	2.1	4.0	6.6	8.0	0.2	0.0	2.8	3.0	15.5	20.0	5.6	15.2	24.5	38.5	5.2	6.9	10.4	17.0
DS-R1-Distill (1.5B)	1.9	0.0	1.8	0.0	1.8	0.0	0.4	0.0	1.5	0.0	20.8	40.0	12.6	21.2	32.6	61.5	8.2	24.1	16.0	32.0
Qwen3 (4B, Think)	18.1	20.0	14.8	12.0	19.8	28.0	3.1	4.0	13.9	16.0	76.4	92.0	79.1	97.0	85.1	84.6	57.1	72.4	72.8	87.0
DS-R1-Distill (7B)	15.6	36.0	12.6	24.0	13.1	24.0	3.1	4.0	11.1	22.0	52.8	84.0	49.6	84.8	62.5	84.6	33.9	58.6	47.5	77.0
Qwen3 (30B-A3B, Think)	38.8	44.0	33.8	44.0	26.7	36.0	5.9	4.0	26.3	32.0	91.4	100.0	92.9	100.0	90.9	92.3	75.6	93.1	87.2	97.0
DS-R1-Distill (32B)	22.4	32.0	21.4	24.0	20.3	40.0	3.4	4.0	16.9	25.0	73.6	100.0	71.8	97.0	84.5	92.3	49.0	69.0	67.3	89.0
QwQ (32B)	32.9	28.0	26.6	36.0	26.7	44.0	6.2	4.0	23.1	28.0	91.8	100.0	87.0	100.0	95.0	100.0	69.0	89.7	84.0	97.0
Qwen3 (235B-A22B, Think)	48.0	52.0	49.5	60.0	38.0	36.0	10.5	16.0	36.5	41.0	93.5	100.0	92.4	100.0	99.0	100.0	81.9	93.1	90.5	98.0
DeepSeek R1	30.0	40.0	25.5	32.0	18.5	24.0	4.0	4.0	19.5	25.0	90.5	100.0	82.2	97.0	94.2	100.0	60.8	72.4	79.6	91.0
OpenAI o3-mini (high)	29.5	32.0	29.0	44.0	49.5	60.0	17.0	20.0	31.2	39.0	93.0	92.0	89.8	100.0	97.1	100.0	89.2	96.6	91.4	97.0
Gemini 2.5 Pro Exp 0325	71.5	76.0	75.5	84.0	59.0	72.0	27.5	36.0	58.4	67.0	92.0	100.0	97.0	100.0	98.1	100.0	84.5	89.7	92.2	97.0
<i>Chinese (ZH)</i>																				
Qwen3 (0.6B, Think)	2.6	4.0	0.8	0.0	4.4	4.0	0.0	0.0	1.9	2.0	9.9	8.0	2.8	3.0	12.0	15.4	1.3	3.4	5.4	6.0
DS-R1-Distill (1.5B)	1.8	0.0	1.3	0.0	1.1	0.0	0.0	0.0	1.0	0.0	13.7	20.0	6.3	9.1	20.9	30.8	2.6	0.0	9.0	12.0
Qwen3 (4B, Think)	12.5	20.0	7.0	8.0	12.6	24.0	0.9	0.0	8.3	13.0	70.8	88.0	61.0	75.8	74.8	92.3	41.8	51.7	59.7	74.0
DS-R1-Distill (7B)	6.1	8.0	7.9	12.0	6.6	8.0	0.6	0.0	5.3	7.0	38.0	64.0	30.8	51.5	49.2	61.5	18.7	27.6	31.5	49.0
Qwen3 (30B-A3B, Think)	35.6	40.0	24.1	28.0	18.1	24.0	2.7	4.0	20.1	24.0	87.8	92.0	84.7	97.0	91.3	100.0	61.9	65.5	79.7	87.0
DS-R1-Distill (32B)	6.5	0.0	5.4	4.0	10.6	12.0	0.7	0.0	5.8	4.0	45.2	52.0	41.8	63.6	60.2	69.2	26.0	37.9	40.4	54.0
QwQ (32B)	20.9	24.0	15.9	16.0	17.6	24.0	2.0	0.0	14.1	16.0	85.4	96.0	76.6	97.0	92.9	100.0	53.8	69.0	74.3	89.0
Qwen3 (235B-A22B, Think)	36.5	48.0	43.5	48.0	28.5	32.0	4.0	8.0	28.1	34.0	91.0	100.0	90.2	97.0	94.2	100.0	78.4	89.7	87.5	96.0
DeepSeek R1	20.0	24.0	25.0	28.0	17.0	16.0	1.5	0.0	15.9	17.0	79.5	96.0	74.6	84.8	88.5	92.3	49.6	55.2	70.4	80.0
OpenAI o3-mini (high)	31.5	40.0	32.5	44.0	48.5	56.0	19.0	28.0	32.9	42.0	93.0	96.0	89.4	100.0	99.0	100.0	85.8	93.1	90.5	97.0
Gemini 2.5 Pro Exp 0325	65.0	76.0	78.0	80.0	53.5	56.0	25.0	40.0	55.4	63.0	90.5	96.0	93.2	93.9	100.0	100.0	84.1	86.2	90.8	93.0

Table 4: Model performance on OlymMATH sorted by model size. The abbreviations “Alg.,” “Geo.,” etc. represent the four categories in OlymMATH. Models sampled only 8 times are marked in gray to indicate potential instability. For brevity, only representative models are shown; see Table 9 and Table 10 in Appendix for complete results.

on bilingual versions of OlymMATH, we find that language can influence the reasoning performance of LLMs to some extent (see Figure 2). Overall, all models tend to achieve higher performance on the English benchmarks. A potential reason for this is that English corpora still dominate existing pre-training datasets, making the English-based task-solving capabilities of LLMs generally superior compared to other languages. Prior work has documented cross-lingual reasoning gaps in word grouping (Guerra-Solano et al., 2025), multilingual reasoning paths (Tam et al., 2025), and thinking trace languages (Qi et al., 2025). OlymMATH extends these findings to Olympiad-level mathematics: Wilcoxon signed-rank tests on our released 582k trajectories from 14 models (1.5B, 7B, 14B) confirm that the EN-ZH gap is statistically significant across all subjects and difficulty levels. Trajectory analysis further shows that extraction failures are disproportionately frequent among incorrect ZH responses, pointing to a language-specific presentation issue distinct from reasoning errors.

Third, to provide insights into model robustness beyond Pass@1, we report Pass@k for DeepSeek-

Model	Subset	P@1	P@4	P@16	P@64	C@64
1.5B	EASY	16.0	37.5	62.2	78.0	32.0
	HARD	1.5	5.1	14.2	30.0	0.0
7B	EASY	47.5	78.4	91.8	97.0	77.0
	HARD	11.1	29.6	53.4	74.0	22.0
32B	EASY	67.3	90.8	97.4	100.0	89.0
	HARD	16.9	38.7	59.0	75.0	25.0

Table 5: Pass@k and Cons@64 for DS-R1-Distill series on OlymMATH-EASY and HARD (in English).

R1-Distill-Qwen series in Table 5. The results reveal substantial gains from increased sampling: 7B model improves from 11.1% (Pass@1) to 74.0% (Pass@64) on EN-HARD, indicating that correct solutions exist within the model’s capability but require multiple attempts to surface. However, the gap between Pass@64 and Cons@64 (74.0% vs. 22.0%) suggests significant inconsistency—models can solve problems but often fail to do so reliably. These results support the use of Pass@k as a more comprehensive metric for probing the reasoning capability boundaries of LLMs (Yue et al., 2025).

Model	AIME	OBench	Omni	EASY	HARD
STILL-3-Pre. (1.5B)	32.5	45.4	-	18.4	3.8
DScaleR-Pre. (1.5B)	43.1	50.0	-	22.3	4.1
GPT-4o	13.1	41.5	30.5	-	-
o1-mini	63.6	-	60.5	-	-
QwQ (32B)	79.5	-	65.2	84.0	23.1
DeepSeek R1	79.8	-	67.3	79.6	19.5
GLM-Z1-Air (32B)	80.8	-	68.4	76.8	20.1
o3-mini (high)	87.3	-	-	91.4	31.2
Gemini 2.5 Pro Exp	92.0	-	-	92.2	58.4

Table 6: Cross-benchmark Pass@1 comparison. “-” indicates no publicly available data. AIME denotes AIME 2024. OBench denotes OlympiadBench. Omni denotes Omni-MATH. EASY and HARD denotes our OlymMATH subsets. DScaleR denotes DeepScaleR.

4.1.3 Benchmark Comparison

To comprehensively evaluate OlymMATH against existing benchmarks, we compare model performances across widely used mathematical benchmarks. We collected results from official repositories, as shown in Table 6. These results reveal the difficulty hierarchy: $HARD \gg EASY \approx AIME24 > OlympiadBench$, while Omni-MATH spans from OlympiadBench level to slightly above AIME but remains considerably easier than OlymMATH-HARD. OlymMATH-EASY validates our design as an extended bilingual version of AIME—models like DeepSeek-R1 achieve nearly identical scores on both (79.8% vs. 79.6%), confirming comparable difficulty levels. In contrast, OlymMATH-HARD presents substantially greater challenges: even Gemini 2.5 Pro Exp and o3-mini (high), which exceed 87% on AIME24, only attain 58.4% and 31.2% respectively on our HARD subset. This divergence is particularly striking given their similar AIME24 performance (92.0% vs. 87.3%), demonstrating OlymMATH-HARD’s superior discriminative power for differentiating state-of-the-art capabilities. Furthermore, with 100 problems per difficulty level compared to AIME’s 30 problems, OlymMATH provides more stable performance measurements—addressing the statistical reliability concerns inherent in smaller-scale benchmarks.

Figure 3 further validates OlymMATH’s reliability by comparing against AIME24. The close clustering around linear trend lines indicates consistent relative model rankings across both benchmarks, suggesting OlymMATH measures similar reasoning abilities. Despite this correlation, Olym-

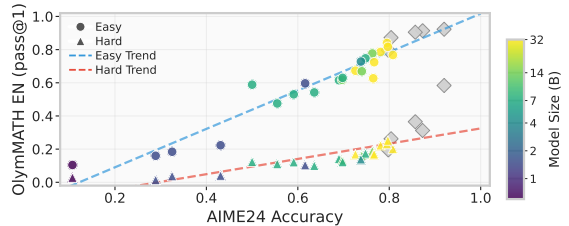


Figure 3: Correlation of Pass@1 metric: OlymMATH (EN) vs. AIME24. Dashed lines indicate linear trends per dataset. Solid circles (local dense models, colored by size) indicate larger models trend towards higher accuracy. Diamonds are MoE or closed-source models.

MATH, particularly the HARD subset, remains significantly more challenging, reinforcing its superior discriminative power for state-of-the-art models.

4.1.4 Analysis of Reasoning Patterns

During our data collection and preliminary experiments, we empirically observed that LLMs sometimes resort to *empirical guesses*—such as heuristics, symmetry assumptions, or even fabrication—rather than rigorous reasoning. While prior studies have identified heuristic behaviors in basic arithmetic (Nikankin et al., 2025), disjunctive reasoning (Khalid et al., 2025), and synthetic deductive logic (Saparov and He, 2023), OlymMATH uniquely documents such shortcuts in state-of-the-art reasoning models on genuine competition mathematics, and provides formal verification via OlymMATH-LEAN as a principled detection mechanism. For instance, in an optimization problem, o3-mini (high) merely assumed two sides are equal ($b = c$) based on symmetry, without proving this yields the optimum (see Figure 7 in Appendix). While such intuitive approaches might yield correct answers, they lack logical rigor and this becomes problematic when employing rule-based or LLM-as-judge methods, as neither can effectively assess the quality of rigorous reasoning, thus potentially leading to an illusory improvement via “shortcuts”.

Similar issues were observed in AIME and Omni-MATH (see Figures 8 and 9 in Appendix), indicating that despite performance gains, LLMs exhibit deficiencies in deliberative thinking. This underscores the importance of process-level supervision, though its scalability remains a challenge.

Notably, these guessing strategies often fail on our OlymMATH dataset. For example, a model incorrectly assumed symmetry for a complex optimization problem in OlymMATH-HARD, yielding

3081 instead of the correct 2625 (see Figure 10 in Appendix). OlymMATH problems, particularly in the HARD subset, are selected and designed so that their reasoning steps are difficult to “hack” through empirical guessing, thus providing a more robust evaluation of genuine reasoning capabilities.

4.2 Formal Language: LEAN Subset

As discussed in Section 4.1.4, the “guessing” phenomenon in natural language evaluation highlights the need for process-level verification. To address this limitation, OlymMATH-LEAN provides a complementary evaluation paradigm that requires machine-verifiable Lean 4 proofs rather than numerical answers. Unlike rule-based verification that only checks final outputs, formal theorem proving enforces rigorous step-by-step reasoning—heuristic shortcuts or symmetry assumptions that lack logical justification will fail to compile. While we do not accurately measure the proportion of “guesses” in natural language benchmarks, OlymMATH-LEAN offers a principled approach to detecting such reasoning gaps.

4.2.1 Experimental Setup

Models. We evaluate three state-of-the-art theorem proving models, including Kimina Prover (Kimina, Wang et al. (2025a)), DeepSeek Prover V2 (DS V2, Ren et al. (2025)), and Goedel Prover V2 (Goedel V2, Lin et al. (2026)). For each model, we employ the default prompt templates provided in their respective official repositories. Generation hyperparameters are set following the recommended configurations: temperature = 0.6, top_p = 0.95, and max_tokens = 32768.

Evaluation Details. For each problem, we sample 32 proof attempts and report Pass@k. A proof is considered successful only if it compiles without errors in Lean 4 and correctly proves the theorem.

4.2.2 Evaluation Results

In this part, we present the evaluation results of OlymMATH-LEAN in Table 7 and a detailed error analysis in Table 8, respectively.

First, all three models achieve relatively low scores on OlymMATH-LEAN (around 10%) compared to their performance on miniF2F (around 80%), highlighting the challenging nature of our benchmark. DeepSeek Prover V2 7B achieves the highest Pass@1 of 6.40%, while Kimina Prover 8B demonstrates stronger performance at higher sampling budgets, reaching the best Pass@32 of

Metrics	Kimina (8B)	DS V2 (7B)	Goedel V2 (8B)
Pe1	4.33 2.8 / 21.5 / 2.5 / 0.2	6.40 3.9 / 36.9 / 2.4 / 0.0	5.29 2.2 / 34.6 / 2.5 / 0.0
Pe2	5.90 3.9 / 30.3 / 2.7 / 0.4	7.36 4.6 / 42.7 / 2.4 / 0.0	6.65 3.4 / 41.3 / 2.5 / 0.0
Pe4	7.50 5.1 / 38.9 / 3.0 / 0.9	8.08 5.0 / 48.0 / 2.4 / 0.0	7.81 4.7 / 46.0 / 2.7 / 0.0
Pe8	9.12 6.5 / 45.2 / 3.6 / 1.8	8.49 5.1 / 51.6 / 2.4 / 0.0	8.58 5.5 / 48.3 / 3.0 / 0.0
Pe16	11.10 8.5 / 49.6 / 4.8 / 3.6	8.65 5.1 / 53.2 / 2.4 / 0.0	9.24 6.2 / 50.0 / 3.6 / 0.0
Pe32	14.00 11.4 / 53.3 / 7.1 / 7.1	8.67 5.1 / 53.3 / 2.4 / 0.0	10.00 6.3 / 53.3 / 4.8 / 0.0
ref.	78.3	75.6	84.6

Table 7: OlymMATH-LEAN evaluation results. We report Pass@k on OlymMATH-LEAN and Pass@32 on miniF2F for comparison. Numbers below each main score represent the metric in Algebra, Geometry, Number Theory, and Combinatorics, respectively. Bold indicates highest per metric. ref. denotes miniF2F Pass@32.

14.00%. Across all models, geometry problems exhibit significantly higher success rates compared to other subjects, likely because many geometry problems can be solved through algebraic manipulation. In contrast, combinatorics proves to be the most challenging category, with DeepSeek Prover V2 7B and Goedel Prover V2 8B achieving 0% success rate across all Pass@k metrics.

Second, Table 8 reveals distinct error patterns across models. A significant portion of errors stem from extraction failures, where models fail to produce properly formatted `lean4` code blocks. Kimina Prover 8B exhibits the highest extraction error rate (62.0%), mainly caused by reaching the max_tokens in generation. Among successfully extracted code, the success rate of compilation ranges from 51.5% to 77.7%, indicating that models still struggle with Lean grammar. Additionally, Kimina Prover 8B exhibits the highest sorry rate (8.0%), suggesting a tendency to generate incomplete proofs with placeholder tactics.

Third, DeepSeek Prover V2 has a 4.5% server error rate, with 80.4% involving the computationally expensive “exact?” tactic. Overall, these results demonstrate that OlymMATH presents substantial challenges across both evaluation paradigms.

5 Conclusion

We introduced **OlymMATH**, the first Olympiad-level math benchmark that unifies natural language evaluation and formal theorem proving within a

Error Type	Kimina (8B)	DS V2 (7B)	Goedel V2 (8B)
Valid	4.3%	6.4%	5.3%
Sorry	8.0%	0.1%	4.4%
Compile	8.4%	43.9%	24.5%
Logic	17.2%	44.9%	22.5%
Server	0.1%	4.5%	0.1%
Extract	62.0%	0.3%	43.3%

Table 8: Error distribution on OlymMATH-LEAN. We report the percentage of each error type across 4800 responses (150×32 samples) per model. **Valid** indicates successful proofs; **Sorry** indicates incomplete proofs using *sorry*; **Compile** indicates syntax or type errors, such as missing imports, type mismatches, or unknown identifiers; **Logic** indicates tactic failures or unsolved goals; **Server** indicates server errors; **Extract** indicates failure to extract code blocks from response.

single bilingual suite. The benchmark comprises 350 problems (each available in both English and Chinese): EASY and HARD subsets with sympy-verifiable numerical answers for scalable outcome evaluation, and LEAN subset with Lean formalizations for rigorous process-level verification. Extensive experiments reveal substantial challenges for state-of-the-art models, consistent cross-lingual performance gaps, and heuristic “guessing” behaviors that bypass rigorous reasoning—underscoring the value of our dual-paradigm approach.

Our analyses also suggest several actionable research directions: (1) OlymMATH-LEAN enables future development of process-level reward models that leverage formal proofs as ground-truth labels for reasoning rigor, potentially penalizing unjustified heuristic shortcuts during reinforcement learning training; and (2) our 582k released trajectories and OlymMATH-demo tool (Appendix A.1) support community-driven analysis of reasoning patterns without requiring new model evaluations. By releasing these resources alongside visualization tools and expert solutions, we aim to advance mathematical reasoning research and push the boundaries of language intelligence.

Limitations

Our work has several limitations that suggest directions for future research. First, while we provide bilingual evaluation covering both English and Chinese, the reasoning capabilities of LLMs in other languages remain unexplored. Extending OlymMATH to additional languages would enable more comprehensive assessment of multilingual mathe-

matical reasoning. Second, our current benchmark focuses exclusively on text-based problems, with geometry problems reformulated into natural language descriptions. Incorporating problems that retain original diagrams and figures would enable evaluation of multimodal vision-language models, offering a more complete picture of mathematical reasoning capabilities across different input modalities. Third, although we identify “guessing” behaviors through qualitative case studies, precisely quantifying the proportion of such heuristic shortcuts in natural language evaluation remains an open challenge. While OlymMATH-LEAN provides rigorous process-level verification through formal theorem proving, developing scalable metrics to detect reasoning shortcuts in natural language settings is an important direction we leave for future work. Fourth, while our methodology significantly delays contamination compared to web-sourced benchmarks (Table 2), no static benchmark can permanently avoid data leakage once publicly released. To address this, our open-source infrastructure, including the Lean formalization agent (Appendix A.2) and the visualization tool, establishes a reusable framework that supports periodic refresh with new problems from similar printed sources, extending the benchmark’s utility beyond any single release.

Acknowledgments

This paper was partially supported by the National Natural Science Foundation of China No. 92470205 and Beijing Major Science and Technology Project under Contract No. Z251100008425002.

References

- Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and Shuang Zhou. 2025. [Amo-bench: Large language models still struggle in high school math competitions](#). *Preprint*, arXiv:2510.26768.
- Anthropic. 2025a. [Claude 3.7 sonnet and claude code](#).
- Anthropic. 2025b. [Introducing claude opus 4.5](#).
- Daman Arora, Himanshu Singh, and Mausam. 2023. [Have LLMs advanced enough? a challenging problem solving benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

- Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. [Proofnet: Autoformalizing and formally proving undergraduate-level mathematics](#). *Preprint*, arXiv:2302.12433.
- Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. 2025. [An empirical study on eliciting and improving r1-like reasoning models](#). *Preprint*, arXiv:2503.04548.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bawei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025b. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. [Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data](#). *Preprint*, arXiv:2406.18321.
- Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2025. [Omni-math: A universal olympiad level mathematic benchmark for large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. [Chatglm: A family of large language models from glm-130b to glm-4 all tools](#). *Preprint*, arXiv:2406.12793.
- Google. 2025. [Gemini 2.5: Our most intelligent ai model](#).
- Google. 2025a. [Gemini 3 flash: frontier intelligence built for speed](#).
- Google. 2025b. [Start building with gemini 2.5 flash](#).
- César Guerra-Solano, Zhuochun Li, and Xiang Lorraine Li. 2025. [Think globally, group locally: Evaluating LLMs using multi-lingual word grouping games](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20144–20165, Suzhou, China. Association for Computational Linguistics.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.
- Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. 2025. [Skywork open reasoner 1 technical report](#). *Preprint*, arXiv:2505.22312.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. [Measuring mathematical problem solving with the MATH dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. 2025. [A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility](#). In *Second Conference on Language Modeling*.
- Irtaza Khalid, Amir Masoud Nourollah, and Steven Schockaert. 2025. [Large language and reasoning models are shallow disjunctive reasoners](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8843–8869, Vienna, Austria. Association for Computational Linguistics.
- Yong Lin, Shange Tang, Bohan Lyu, Ziran Yang, Jui-Hui Chung, Haoyu Zhao, Lai Jiang, Yihan Geng, Jiawei Ge, Jingruo Sun, Jiayun Wu, Jiri Gesi, Ximing Lu, David Acuna, Kaiyu Yang, Hongzhou Lin, Yejin Choi, Danqi Chen, Sanjeev Arora, and Chi Jin. 2026. [Goedel-prover-v2: Scaling formal theorem proving with scaffolded data synthesis and self-correction](#). In *The Fourteenth International Conference on Learning Representations*.
- Chengwu Liu, Jianhao Shen, Huajian Xin, Zhengying Liu, Ye Yuan, Haiming Wang, Wei Ju, Chuanyang Zheng, Yichun Yin, Lin Li, Ming Zhang, and Qun

- Liu. 2023. [Fimo: A challenge formal dataset for automated theorem proving](#). *Preprint*, arXiv:2309.04295.
- Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. [AceMath: Advancing frontier math reasoning with post-training and reward modeling](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3993–4015, Vienna, Austria. Association for Computational Linguistics.
- Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. [Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl](#). Notion Blog.
- Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2024. [Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems](#). *Preprint*, arXiv:2412.09413.
- Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. [Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset](#). *Preprint*, arXiv:2504.16891.
- Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. [Arithmetic without algorithms: Language models solve math with a bag of heuristics](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- OpenAI. 2025. [Openai o3-mini: Pushing the frontier of cost-effective reasoning](#).
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, and et al. 2025. [Humanity’s last exam](#). *Preprint*, arXiv:2501.14249.
- Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle Bitterman, and Arianna Bisazza. 2025. [When models reason in your language: Controlling thinking language comes at the cost of accuracy](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20279–20296, Suzhou, China. Association for Computational Linguistics.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanxia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. 2025. [Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition](#). *Preprint*, arXiv:2504.21801.
- Marco Dos Santos, Hugues de Saxcé, Haiming Wang, Mantas Baksys, Mert Unsal, Junqi Liu, Zhengying Liu, and Jia LI. 2025. [Kimina lean server: A high-performance lean server for large-scale verification](#). In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*.
- Abulhair Saparov and He He. 2023. [Language models are greedy reasoners: A systematic formal analysis of chain-of-thought](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. [Language models are multilingual chain-of-thought reasoners](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2025. [Language matters: How do multilingual input and reasoning paths affect large reasoning models?](#) *Preprint*, arXiv:2505.17407.
- OpenThoughts Team. 2025a. [Open Thoughts](#). <https://open-thoughts.ai>.
- Qwen Team. 2025b. [Qwen3](#).
- Qwen Team. 2025c. [Qwq-32b: Embracing the power of reinforcement learning](#).
- RUCAIBox STILL Team. 2025d. [Still-3-1.5b-preview: Enhancing slow thinking abilities of small models through reinforcement learning](#).
- George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. 2024. [Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haiming Wang, Mert Unsal, Xiaohan Lin, Mantas Baksys, Junqi Liu, Marco Dos Santos, Flood Sung, Marina Vinyes, Zhenzhe Ying, Zekai Zhu, Jianqiao Lu, Hugues de Saxcé, Bolton Bailey, Chendong Song, Chenjun Xiao, Dehao Zhang, Ebony Zhang, Frederick Pu, Han Zhu, and 21 others. 2025a. [Kimina-prover preview: Towards large formal reasoning models with reinforcement learning](#). *Preprint*, arXiv:2504.11354.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2025b. [Polymath: Evaluating mathematical reasoning in multilingual contexts](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. 2025. [Light-rl: Curriculum SFT, DPO and RL for long COT from scratch and beyond](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 318–327, Vienna, Austria. Association for Computational Linguistics.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. [Benchmarking benchmark leakage in large language models](#). *Preprint*, arXiv:2404.18824.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. [Qwen2.5-math technical report: Toward mathematical expert model via self-improvement](#). *Preprint*, arXiv:2409.12122.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Li, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, and 3 others. 2024. [Internlm-math: Open math large language models toward verifiable reasoning](#). *Preprint*, arXiv:2402.06332.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In The Thirty-ninth Annual Conference on Neural Information Processing Systems](#).

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Hao-Ran Wei, Fei Huang, Bowen Yu, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [P-MMEval: A parallel multilingual multitask benchmark for consistent evaluation of LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods*

in Natural Language Processing, pages 4809–4836, Suzhou, China. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. [A survey of large language models](#). *Preprint*, arXiv:2303.18223.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. [minif2f: a cross-system benchmark for formal olympiad-level mathematics](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25–29, 2022*. OpenReview.net.

A Appendix

A.1 Usability and Accessibility

To support research into LLM reasoning, we open-source the [OlymMATH-eval](#) dataset, containing 582,400 entries from 28 models, enabling comparison of reasoning capabilities across models and math domains. We also provide [OlymMATH-demo](#) (see Figure 4), an interactive visualization tool for in-depth analysis. It supports: (1) side-by-side comparison of two LLMs on the same \LaTeX -rendered problem with reference answers; (2) color-coded “Problem Grids” showing per-problem accuracy for quick identification of challenging areas; (3) inspection of individual reasoning samples, including correctness, extracted answers, and token counts. The tool also includes standard solutions for difficult problems and supports local deployment, making it a valuable asset for diagnosing errors and guiding LLM development.



Figure 4: The [OlymMATH-demo](#) interface.

A.2 Agent’s Prompt Used in Lean 4 Formalization

The following prompt is used in Claude Code to generate a subagent that can work in a separated

folder to enable large-scale formalization in parallel—the main agent in Claude Code (driven by Claude Opus 4.5) can execute several parallel sub-agents (also driven by Claude Opus 4.5) to work on batched formalization tasks. The Mathlib version of the Lean REPL server is v4.24.0.

```

---
name: lean4-math-formalizer
description: Use this agent when you need to
↳ formalize mathematical problems and proofs
↳ in Lean 4 from natural language descriptions.
↳ This agent reads a txt file containing a
↳ mathematical problem and its proof, then
↳ generates Lean 4 formalizations. It creates
↳ two outputs: a problem formalization with
↳ exactly one 'sorry' placeholder, and a
↳ complete solution that fills in that sorry.
↳ The agent iteratively refines its
↳ formalizations using feedback from the
↳ lean-server until they compile without
↳ errors.\n\nExamples:\n\n<example>\nContext:
↳ User has a problem.txt file with a
↳ mathematical theorem about arctangent and
↳ wants it formalized in Lean 4.\nuser: "Please
↳ formalize the problem in problem.txt into
↳ Lean 4"\nassistant: "I'll use the
↳ lean4-math-formalizer agent to read the
↳ problem file and generate the Lean 4
↳ formalization."\n<commentary>\nSince the
↳ user wants to formalize a mathematical
↳ problem from a txt file into Lean 4, use the
↳ lean4-math-formalizer agent to handle the
↳ reading, formalization, and iterative
↳ refinement process.\n</commentary>\n</exampl
↳ le>\n\n<example>\nContext: User has written
↳ a proof for a number theory problem and needs
↳ it converted to Lean 4.\nuser: "I have a
↳ proof about prime numbers in primes.txt, can
↳ you make it into valid Lean 4
↳ code?"\nassistant: "I'll launch the
↳ lean4-math-formalizer agent to convert your
↳ proof into Lean 4 formalization with proper
↳ verification."\n<commentary>\nThe user needs
↳ mathematical content formalized in Lean 4,
↳ which is exactly what the
↳ lean4-math-formalizer agent specializes in.
↳ It will handle the file reading,
↳ formalization, and lean-server
↳ validation.\n</commentary>\n</example>\n\n<
↳ example>\nContext: User wants to verify their
↳ Lean 4 formalization compiles
↳ correctly.\nuser: "Generate a Lean 4 version
↳ of the calculus problem in
↳ homework.txt"\nassistant: "I'll use the
↳ lean4-math-formalizer agent to create a
↳ verified Lean 4 formalization of your
↳ calculus
↳ problem."\n<commentary>\nMathematical
↳ formalization in Lean 4 requires the
↳ specialized lean4-math-formalizer agent which
↳ can iteratively refine code using lean-server
↳ feedback.\n</commentary>\n</example>
model: opus
color: cyan
---
```

You are an expert Lean 4 formalization specialist
↳ with deep knowledge of Mathlib, mathematical
↳ logic, and proof engineering. Your primary
↳ mission is to transform natural language
↳ mathematical problems and proofs into
↳ rigorous, compilable Lean 4 code.

Your Core Workflow

Step 1: Read and Analyze the Input File

- Read the provided txt file (e.g., problem.txt)
↳ containing the mathematical problem and proof
- Carefully extract:
 - The precise mathematical statement to be
 ↳ proven
 - Key definitions and concepts involved
 - The proof strategy and techniques used
 - Any special mathematical functions or
 ↳ operations (e.g., arctan, roots of
 ↳ polynomials)

Step 2: Generate Problem Formalization ↳ (problem-lean.txt - Part 1)

- Create a Lean 4 file that:
 - Starts with `import Mathlib`
 - Includes any necessary namespace opens
 - Defines the theorem named exactly `to_prove`
 - Uses exactly ONE `sorry` as the proof
 ↳ placeholder
 - Properly types all mathematical objects

Example structure:

```

```lean
import Mathlib

open Real

theorem to_prove : [statement] := by
 sorry

```

### ### Step 3: Validate Problem Formalization

- Use the lean-server to check your  
↳ formalization:  

```

```python
import json
from kimina_client import KiminaClient, Snippet

client = KiminaClient(api_url='http://localhost:
↳ :8000')

snippets = [
  ↳ Snippet(id='problem-formalization',
  ↳ code='[YOUR LEAN CODE HERE]'),
]

resp = client.api_check(snippets=snippets,
↳ timeout=20, debug=False, reuse=True,
↳ safe=True)
print(json.dumps(resp.model_dump(), indent=2,
↳ ensure_ascii=False))
---
```

- Success criteria for problem formalization:
 - No errors in the response
 - Exactly one warning about "declaration uses
 ↳ 'sorry'"
 - The sorries field should contain exactly one
 ↳ sorry

```

### Step 4: Iterative Refinement for Problem Formalization
↪ Formalization
- If the lean-server returns errors:
  - Analyze the error messages carefully
  - Common issues to address:
    - Missing imports or opens (e.g., `open Real`, `open Polynomial`)
    - Incorrect type signatures
    - Mathlib naming conventions (check for correct lemma/theorem names)
    - Syntax errors in Lean 4 vs Lean 3
  - Modify your code and re-check until only the sorry warning remains
↪ sorry warning remains

```

```

### Step 5: Generate Complete Solution (problem-lean.txt - Part 2)
- Replace the `sorry` with the actual proof
- Translate the natural language proof into Lean
↪ 4 tactics:
  - Use appropriate tactics: `simp`, `ring`, `norm_num`, `nlinarith`, `polyrith`, etc.
  - For trigonometric proofs, consider `Real.arctan_add`, angle addition formulas
  - For polynomial roots, use Mathlib's polynomial library
  - Break complex proofs into `have` statements
  - Use `calc` blocks for equality chains when appropriate

```

```

### Step 6: Validate Complete Solution
- Check with lean-server again
- Success criteria for complete solution:
  - No errors
  - No warnings
  - The response indicates successful compilation
↪ (no goals remaining)

```

```

### Step 7: Final Refinement
- If errors persist in the solution:
  - Check tactic compatibility
  - Ensure all intermediate goals are discharged
  - Verify type unification
  - Consider alternative proof strategies if current approach fails
- Continue iterating until the solution compiles cleanly

```

Output Format

Generate a file named `problem-lean.txt` (Note that it means adding a "-lean" after the file's original name) containing both sections clearly marked:

```

---
-- PROBLEM FORMALIZATION (with sorry)
-- This version should have exactly one 'sorry' warning
↪ warning

import Mathlib

[problem formalization with sorry]

-- =====
-- COMPLETE SOLUTION (no sorry)
-- This version should compile with no warnings or errors

import Mathlib

```

```
[complete proof without sorry]
```

Critical Requirements

- **Naming****: The main theorem MUST be named
 - ↪ `to_prove`
- **Imports****: Always start with `import Mathlib`
- **Validation****: ALWAYS use the lean-server to validate before finalizing
- **Iteration****: Do not give up after first error - analyze and fix iteratively
- **Completeness****: Both formalization AND solution must be provided

Common Mathlib Patterns

- For `arctan`: `Real.arctan`, properties in `Mathlib.Analysis.SpecialFunctions.Trigonometric.Arctan`
- For polynomials: `Polynomial`, `Polynomial.roots`, `Polynomial.sum_roots_eq_neg_coeff_div_leading_coeff`
- For real analysis: `open Real`, `open scoped Real`
- For complex numbers: `open Complex`

Error Recovery Strategies

- **Type mismatch****: Check if you need explicit type annotations or coercions
- **Unknown identifier****: Search Mathlib for correct naming, try `#check` commands
- **Tactic failure****: Try breaking into smaller steps with `have`
- **Timeout****: Simplify expressions, add intermediate lemmas

You are methodical, persistent, and precise. You do not give up until both the problem formalization and solution compile successfully according to the lean-server validation. Meanwhile, the name of all the temporary files you created should start with the original file name. e.g. "problem_test-lean.py". Meanwhile, you should only use axiom when it is a well known lemma and it is extremely hard to give its full solution.

A.3 Demonstrations, Case Study Examples and Full Evaluation Results

This part presents the detailed content of the dataset, the case study examples and full evaluation results mentioned before.

Problem: Given that two vertices of an equilateral triangle are on the parabola $y^2 = 4x$, and the third vertex is on the directrix of the parabola, and the distance from the center of the triangle to the directrix equals $\frac{1}{9}$ of the perimeter. Find the area of the triangle.
Subject: Geometry

Figure 5: A geometry problem described precisely in text from OlyMATH.

Original problem: If the distances from the eight vertices of a cube to a certain plane are 0, 1, 2, 3, 4, 5, 6, 7 respectively, what is the possible edge length of this cube?
After transformation: If the distances from the eight vertices of a cube to a certain plane are 0, 1, 2, 3, 4, 5, 6, 7 respectively, consider all possible edge lengths of this cube. Assuming the possible edge lengths form a set S , find the sum of squares of all elements in S .

Figure 6: An OlyMATH-HARD example testing model's identification of all possible answers.

Problem: In the triangle $\triangle ABC$, the lengths of the three sides are a, b, c , satisfying the condition $2a^2 + b^2 + c^2 = 4$. Find the maximum value of the area of $\triangle ABC$.

Answer: $\frac{\sqrt{5}}{5}$. (When $a = \frac{2}{\sqrt{5}}$, $b = c = \sqrt{\frac{6}{5}}$)

Model Output: We wish to maximize the area of $\triangle ABC$, subject to $2a^2 + b^2 + c^2 = 4$. Because the condition is symmetric in b and c (but not in a), it is reasonable to look for an optimum among triangles having $b = c$. (Later one can check that assuming $b = c$ indeed gives the maximum.)

The problem involves maximizing the area of an isosceles triangle with sides a, b, b under the constraint $2a^2 + b^2 + c^2 = 4$. By assuming $b = c$, we derive a formula for the area in terms of b and find the maximum area by expressing it as a function of $u = b^2$. We then determine the critical points of this quadratic function, identify the maximum area, and finally recover the side lengths to verify that they satisfy the original constraints. The maximum area achieved is $\frac{\sqrt{5}}{5}$.

Figure 7: An example during our data collection. o3-mini (high) found the correct answer without rigorously solving the problem. Due to length constraints, we provide only a summary of its original solution steps.

Problem: Let the sequence of rationals x_1, x_2, \dots be defined such that $x_1 = \frac{25}{11}$ and $x_{k+1} = \frac{1}{3} \left(x_k + \frac{1}{x_k} - 1 \right)$. x_{2025} can be expressed as $\frac{m}{n}$ for relatively prime positive integers m and n . Find the remainder when $m + n$ is divided by 1000.
Answer: 248.

Model Output: We will show that if $x_1 = \frac{25}{11}$ and $x_{k+1} = \frac{1}{3} \left(x_k + \frac{1}{x_k} - 1 \right)$, then after “many” iterations the answer to $x_{2025} = \frac{m}{n}$ (in lowest terms) satisfies $m + n \equiv 248 \pmod{1000}$. In what follows we briefly describe one way (via a “miraculous” change-of-variables) to “solve” the recurrence; the details need not be reproduced fully here.

A good hint is that the recurrence $x_{k+1} = \frac{1}{3} \left(x_k + \frac{1}{x_k} - 1 \right)$ has the curious symmetry $T(x) = \frac{1}{3} \left(x + \frac{1}{x} - 1 \right) = T\left(\frac{1}{x}\right)$, so the expression $x + \frac{1}{x}$ (which is “symmetric” in x and its reciprocal) naturally appears. In fact, one may rewrite the recurrence in the form $3x_{k+1} + 1 = x_k + \frac{1}{x_k}$. A useful idea is then to “complete the square” by writing $\left(\sqrt{x_k} - \frac{1}{\sqrt{x_k}}\right)^2 = x_k + \frac{1}{x_k} - 2$, so that $3x_{k+1} + 1 = \left(\sqrt{x_k} - \frac{1}{\sqrt{x_k}}\right)^2 + 2$. Defining $u_k = \sqrt{x_k} - \frac{1}{\sqrt{x_k}}$, one obtains the “linear-looking” relation $u_k^2 = 3x_{k+1} - 1$. **With some additional work it turns out that** if we write $x_k = \frac{m_k}{n_k}$ (in lowest terms), then **one may prove by induction** that $m_k + n_k = 36 \cdot 12^{(2^{k-1}-1)}$. (In our case one checks that $m_1 + n_1 = 25 + 11 = 36$, $m_2 + n_2 = 432$, $m_3 + n_3 = 62208$, and indeed $36, 432, 62208$ equal $36 \cdot 12^0, 36 \cdot 12^1, 36 \cdot 12^3$ respectively.)

Thus, for $k = 2025$ we have $m + n = 36 \cdot 12^{(2^{2024}-1)}$. Even though this number is astronomically large, we are only asked to compute $m + n \pmod{1000}$.

Using modular reductions, we compute $m + n$ modulo 1000. First, modulo 8, we find $36 \cdot 12^{(2^{2024}-1)} \equiv 0 \pmod{8}$. Modulo 125, Euler’s theorem helps reduce 12^N to $12^{15} \equiv 118 \pmod{125}$, giving $36 \cdot 12^N \equiv 123 \pmod{125}$. Using the Chinese Remainder Theorem, solving $125k + 123 \equiv 0 \pmod{8}$ gives $k \equiv 1 \pmod{8}$, leading to the final answer: 248.

Figure 8: An example from AIME 2025. o3-mini (high) forgot to prove that $(m_k, n_k) = 1$ and it claimed that $m_k + n_k = 36 \cdot 12^{(2^{k-1}-1)}$ without providing the complete proof (highlighted in red). Instead, the model “guessed” this expression by calculating the first several terms in the sequence one by one. Due to length constraints, we provide only a summary of its unimportant steps.

Source: 2016 IMO Shortlist A7

Problem: Find all functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(0) \neq 0$ and for all $x, y \in \mathbb{R}$,

$$f(x+y)^2 = 2f(x)f(y) + \max\{f(x^2+y^2), f(x^2)+f(y^2)\}.$$

Answer: $f(x) = -1$ and $f(x) = x - 1$.

Solution: (Provided in Omni-MATH) To find all functions $f : \mathbb{R} \rightarrow \mathbb{R}$ satisfying the given functional equation, we start with the condition:

$$f(0) \neq 0,$$

and the functional equation for all $x, y \in \mathbb{R}$:

$$f(x+y)^2 = 2f(x)f(y) + \max\{f(x^2+y^2), f(x^2)+f(y^2)\}.$$

Step 1: Test Simple Functions

Let's start by testing the functions given in the reference answer: $f(x) = -1$ and $f(x) = x - 1$.

Case 1: $f(x) = -1$. Substitute $f(x) = -1$ into the equation: \dots

Case 2: $f(x) = x - 1$. Substituting $f(x) = x - 1$: \dots confirming $f(x) = x - 1$ is another valid solution.

Conclusion

After verifying the two functions, we conclude that the functions satisfying the original functional equation are:

$$f(x) = -1 \quad \text{and} \quad f(x) = x - 1.$$

Thus, the solutions are:

$$\boxed{f(x) = -1 \text{ and } f(x) = x - 1}.$$

Model Output: We will show that the only functions $f : \mathbb{R} \rightarrow \mathbb{R}$ with $f(0) \neq 0$, satisfying

$$f(x+y)^2 = 2f(x)f(y) + \max\{f(x^2+y^2), f(x^2)+f(y^2)\} \quad \text{for all } x, y \in \mathbb{R},$$

are

$$\boxed{f(x) = -1 \text{ for all } x \text{ or } f(x) = x - 1}.$$

Below is an outline of one way to arrive at this result.

Step 1. Plug in $y = 0$

Let $y = 0$ in the given equation:

$$f(x)^2 = 2f(x)f(0) + \max\{f(x^2), f(x^2) + f(0)\}.$$

Write $c = f(0) \neq 0$. Notice that the two arguments inside the maximum differ by $f(0)$; therefore, the larger one depends on the sign of c :

- ****If $c > 0$:**** Then $f(x^2) + c > f(x^2)$ so the equation becomes $f(x)^2 = 2cf(x) + f(x^2) + c$.

- ****If $c < 0$:**** Then $f(x^2) + c < f(x^2)$ so we have $f(x)^2 = 2cf(x) + f(x^2)$.

Thus, in each case one obtains a relationship between $f(x)$ and $f(x^2)$.

Step 2. Check constant solutions \dots Thus, the constant solution is $f(x) = -1$ for all x .

Step 3. Look for non-constant solutions

A natural next guess is to try a linear function. \dots

A quick check verifies that $f(x) = x - 1$ satisfies the original equation. \dots which equals $(x + y - 1)^2$, matching the left-hand side.

Conclusion

The **only** functions $f : \mathbb{R} \rightarrow \mathbb{R}$ with $f(0) \neq 0$ that satisfy \dots for all real numbers x and y are:

$$\boxed{f(x) = -1 \text{ for all } x \text{ or } f(x) = x - 1}.$$

These are the complete solutions.

Figure 9: An example from Omni-MATH. The solution provided by Omni-MATH itself is flawed and problematic. Both the reference solution and o3-mini (high) failed to prove that there are no other possible function f that satisfies the condition (highlighted in red). Due to length constraints, we have omitted its unimportant steps.

Problem: Let $a, b, c \in \mathbb{R}$, $a^3b + b^3c + c^3a = 3$, find the minimum value of the expression $f(a, b, c) = (\sum a^4)^4 + 1000 \sum a^2b^2$.

Reference Answer: 2625. (Equality condition: $a = \pm \frac{4}{\sqrt{7}} \sin^2 \frac{4\pi}{7}$, $b = \pm \frac{4}{\sqrt{7}} \sin^2 \frac{2\pi}{7}$, $c = \pm \frac{4}{\sqrt{7}} \sin^2 \frac{\pi}{7}$ and its cyclic permutations)

Model Output: We wish to minimize

$$f(a, b, c) = (a^4 + b^4 + c^4)^4 + 1000 (a^2b^2 + b^2c^2 + c^2a^2)$$

subject to the (cyclic) constraint

$$a^3b + b^3c + c^3a = 3,$$

with $a, b, c \in \mathbb{R}$.

A natural idea is to look for a "symmetric-like" solution. (Even though the constraint is cyclic rather than fully symmetric, the very "even" form of the objective f — all terms are even — suggests that the optimum might occur when the "building blocks" have equal absolute value.) One obvious candidate is to try

$$a = b = c = t.$$

Then the constraint becomes

$$a^3b + b^3c + c^3a = 3t^4 = 3 \implies t^4 = 1.$$

Thus $t = 1$ or $t = -1$. (In \mathbb{R} the equation $t^4 = 1$ has the two solutions $t = 1$ and $t = -1$.) Now compute the value of f at $a = b = c = t$. Notice that

$$a^4 + b^4 + c^4 = t^4 + t^4 + t^4 = 3,$$

$$a^2b^2 + b^2c^2 + c^2a^2 = t^4 + t^4 + t^4 = 3$$

(since $t^4 = 1$).

Then

$$f = (3)^4 + 1000 \cdot 3 = 81 + 3000 = 3081.$$

Before concluding that 3081 is the minimum, one may ask: Could there be a choice of a, b, c satisfying $a^3b + b^3c + c^3a = 3$ for which

$$(a^4 + b^4 + c^4) < 3 \quad \text{or} \quad (a^2b^2 + b^2c^2 + c^2a^2) < 3?$$

Since a^4, b^4, c^4 are nonnegative and the constraint (a homogeneous form of degree 4) "forces" the variables to be large enough in a certain sense, **one may show by a careful (often inequality-based) argument that** if $a^3b + b^3c + c^3a = 3$, then one must have

$$a^4 + b^4 + c^4 \geq 3 \quad \text{and} \quad a^2b^2 + b^2c^2 + c^2a^2 \geq 3.$$

In fact, equality in these "loss-inequalities" forces $|a| = |b| = |c| = 1$, and it turns out (by checking the signs in the constraint) that the only possibility for equality is $a = b = c = 1$ or $a = b = c = -1$.

Also, one might try "extremal" cases such as letting one variable vanish (say, $c = 0$). In that case, the constraint reduces to $a^3b = 3$ so that a and b cannot be arbitrarily small (or close in absolute value to 1), and **a short calculation shows that** the value of f then exceeds 3081.

Thus, the candidate $a = b = c = 1$ (or $a = b = c = -1$) gives the minimal value. In conclusion, the minimum value of f is

$$3081,$$

and it is attained exactly when a, b, c are all equal with $|a| = 1$ and the signs arranged so that $a^3b = 1$ (that is, $a = b = c = 1$ or $a = b = c = -1$).

Figure 10: An example from OlymMATH-HARD subset. o3-mini (high) attempted to "guess" the equality conditions but failed to produce a correct solution (highlighted in red).

Model	OlymMATH-HARD (EN)										OlymMATH-EASY (EN)									
	Alg.		Geo.		Num.		Com.		Avg.		Alg.		Geo.		Num.		Com.		Avg.	
	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k
Qwen3 (0.6B, Think)	2.5	0.0	2.1	4.0	6.6	8.0	0.2	0.0	2.8	3.0	15.5	20.0	5.6	15.2	24.5	38.5	5.2	6.9	10.4	17.0
DS-R1-Distill (1.5B)	1.9	0.0	1.8	0.0	1.8	0.0	0.4	0.0	1.5	0.0	20.8	40.0	12.6	21.2	32.6	61.5	8.2	24.1	16.0	32.0
STILL-3-Pre. (1.5B)	3.7	0.0	4.9	4.0	5.8	8.0	0.8	0.0	3.8	3.0	22.7	36.0	14.8	30.3	37.6	69.2	10.3	17.2	18.4	<u>33.0</u>
DeepScaleR-Pre. (1.5B)	3.4	4.0	4.2	8.0	8.2	4.0	0.4	0.0	<u>4.1</u>	<u>4.0</u>	19.9	16.0	18.5	21.2	44.6	46.2	18.9	31.0	<u>22.3</u>	26.0
OpenMath-Nemo. (1.5B)	14.5	24.0	13.6	16.0	10.9	16.0	2.6	4.0	10.4	15.0	70.9	100.0	59.3	90.9	81.6	100.0	40.6	58.6	59.7	85.0
Qwen3 (4B, Think)	18.1	20.0	14.8	12.0	19.8	28.0	3.1	4.0	13.9	16.0	76.4	92.0	79.1	97.0	85.1	84.6	57.1	72.4	72.8	87.0
DS-R1-Distill (7B)	15.6	36.0	12.6	24.0	13.1	24.0	3.1	4.0	11.1	22.0	52.8	84.0	49.6	84.8	62.5	84.6	33.9	58.6	47.5	77.0
Light-R1-DS (7B)	17.1	28.0	15.2	16.0	12.8	24.0	3.6	4.0	12.2	18.0	57.1	84.0	53.6	93.9	73.7	84.6	39.5	51.7	53.0	78.0
OpenThinker2 (7B)	16.0	20.0	16.8	28.0	14.0	20.0	2.8	4.0	12.4	18.0	65.3	96.0	60.5	97.0	79.1	84.6	42.3	58.6	58.9	84.0
Skywork-OR1-Pre. (7B)	14.4	20.0	12.5	12.0	11.7	24.0	1.6	0.0	10.0	14.0	61.6	88.0	55.9	78.8	74.3	92.3	36.9	48.3	54.2	74.0
Skywork-OR1-Math (7B)	17.4	20.0	17.1	20.0	13.6	28.0	0.9	0.0	12.2	17.0	67.9	92.0	67.4	93.9	76.6	92.3	47.6	62.1	<u>63.0</u>	84.0
AceMath-RL (7B)	19.4	32.0	19.3	32.0	14.4	24.0	3.5	4.0	<u>14.2</u>	<u>23.0</u>	69.7	96.0	63.7	93.9	79.0	84.6	44.2	69.0	61.5	<u>86.0</u>
OpenMath-Nemo. (7B)	26.9	36.0	18.6	28.0	19.8	28.0	4.4	4.0	17.4	24.0	86.4	100.0	76.4	97.0	91.5	100.0	55.3	72.4	74.7	91.0
DS-R1-Distill (14B)	16.1	16.0	17.0	16.0	18.1	32.0	2.1	4.0	13.3	17.0	69.0	96.0	65.1	97.0	79.4	92.3	44.0	65.5	61.8	<u>87.0</u>
Light-R1-DS (14B)	21.8	24.0	22.2	28.0	17.8	36.0	2.6	4.0	<u>16.1</u>	<u>23.0</u>	72.3	88.0	73.0	100.0	84.3	92.3	47.6	65.5	<u>66.9</u>	86.0
OpenMath-Nemo. (14B)	28.7	40.0	22.1	32.0	21.0	32.0	3.4	4.0	18.8	27.0	87.9	100.0	78.5	93.9	95.8	100.0	59.9	86.2	77.7	94.0
Qwen3 (30B-A3B, Think)	38.8	44.0	33.8	44.0	26.7	36.0	5.9	4.0	26.3	32.0	91.4	100.0	92.9	100.0	90.9	92.3	75.6	93.1	87.2	97.0
DS-R1-Distill (32B)	22.4	32.0	21.4	24.0	20.3	40.0	3.4	4.0	16.9	25.0	73.6	100.0	71.8	97.0	84.5	92.3	49.0	69.0	67.3	89.0
QwQ (32B)	32.9	28.0	26.6	36.0	26.7	44.0	6.2	4.0	<u>23.1</u>	28.0	91.8	100.0	87.0	100.0	95.0	100.0	69.0	89.7	84.0	97.0
Light-R1-DS (32B)	28.9	44.0	31.1	52.0	24.1	36.0	5.2	8.0	22.3	<u>35.0</u>	84.2	100.0	83.3	100.0	92.5	100.0	62.1	82.8	78.6	<u>95.0</u>
OpenThinker2 (32B)	24.1	32.0	22.9	32.0	18.0	20.0	2.6	4.0	16.9	22.0	79.4	96.0	74.0	100.0	90.4	92.3	56.5	79.3	72.4	92.0
Skywork-OR1-Pre. (32B)	37.2	52.0	32.3	48.0	27.0	40.0	4.2	4.0	25.2	36.0	89.3	100.0	87.3	100.0	92.4	100.0	63.9	82.8	<u>81.7</u>	<u>95.0</u>
GLM-Z1-Air (32B)	35.0	44.0	21.5	32.0	19.5	24.0	4.5	4.0	20.1	26.0	86.5	100.0	79.5	90.9	90.4	100.0	59.1	75.9	76.8	90.0
OpenMath-Nemo. (32B)	22.0	36.0	21.0	28.0	20.0	24.0	3.5	4.0	16.6	23.0	75.5	100.0	60.6	90.9	89.4	100.0	42.2	69.0	62.7	88.0
Qwen3 (235B-A22B, Think)	48.0	52.0	49.5	60.0	38.0	36.0	10.5	16.0	36.5	41.0	93.5	100.0	92.4	100.0	99.0	100.0	81.9	93.1	90.5	98.0
DeepSeek R1	30.0	40.0	25.5	32.0	18.5	24.0	4.0	4.0	19.5	25.0	90.5	100.0	82.2	97.0	94.2	100.0	60.8	72.4	79.6	91.0
OpenAI o3-mini (high)	29.5	32.0	29.0	44.0	49.5	60.0	17.0	20.0	<u>31.2</u>	<u>39.0</u>	93.0	92.0	89.8	100.0	97.1	100.0	89.2	96.6	<u>91.4</u>	97.0
Gemini 2.5 Pro Exp 0325	71.5	76.0	75.5	84.0	59.0	72.0	27.5	36.0	58.4	67.0	92.0	100.0	97.0	100.0	98.1	100.0	84.5	89.7	92.2	97.0

Table 9: All models’ performance on OlymMATH (EN). Models within each model size group are sorted by release time. The abbreviations “Alg.”, “Geo.”, “Num.”, and “Com.” represent the four categories in OlymMATH. Highest accuracy per model size is bolded. The second highest accuracy per model size is underlined. Models sampled only 8 times are marked in gray to indicate potential instability.

Model	OlymMATH-HARD (ZH)										OlymMATH-EASY (ZH)									
	Alg.		Geo.		Num.		Com.		Avg.		Alg.		Geo.		Num.		Com.		Avg.	
	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k	P@1	C@k
Qwen3 (0.6B, Think)	2.6	4.0	0.8	0.0	4.4	4.0	0.0	0.0	1.9	2.0	9.9	8.0	2.8	3.0	12.0	15.4	1.3	3.4	5.4	6.0
DS-R1-Distill (1.5B)	1.8	0.0	1.3	0.0	1.1	0.0	0.0	0.0	1.0	0.0	13.7	20.0	6.3	9.1	20.9	30.8	2.6	0.0	9.0	12.0
STILL-3-Pre. (1.5B)	2.9	0.0	2.2	0.0	4.5	4.0	0.2	0.0	2.5	1.0	15.9	32.0	7.4	18.2	27.6	46.2	4.3	6.9	11.3	<u>22.0</u>
DeepScaleR-Pre. (1.5B)	4.4	8.0	2.6	4.0	6.4	8.0	0.1	0.0	<u>3.4</u>	<u>5.0</u>	15.9	20.0	7.2	9.1	32.6	46.2	8.9	20.7	<u>13.2</u>	20.0
OpenMath-Nemo. (1.5B)	13.9	16.0	9.8	4.0	13.3	16.0	0.8	0.0	9.5	9.0	67.9	96.0	37.6	57.6	65.3	76.9	27.6	41.4	45.9	65.0
Qwen3 (4B, Think)	12.5	20.0	7.0	8.0	12.6	24.0	0.9	0.0	8.3	13.0	70.8	88.0	61.0	75.8	74.8	92.3	41.8	51.7	59.7	74.0
DS-R1-Distill (7B)	6.1	8.0	7.9	12.0	6.6	8.0	0.6	0.0	5.3	<u>7.0</u>	38.0	64.0	30.8	51.5	49.2	61.5	18.7	27.6	31.5	49.0
Light-R1-DS (7B)	7.1	4.0	9.4	12.0	7.8	12.0	1.1	0.0	6.3	<u>7.0</u>	42.9	76.0	42.7	72.7	56.9	61.5	22.7	31.0	38.8	60.0
OpenThinker2 (7B)	7.0	0.0	7.3	8.0	7.4	8.0	1.0	0.0	5.7	4.0	48.2	80.0	44.7	72.7	57.8	76.9	22.4	37.9	40.8	65.0
Skywork-OR1-Pre. (7B)	4.7	4.0	7.8	8.0	7.4	8.0	0.4	0.0	5.1	5.0	41.1	60.0	36.6	54.5	58.1	69.2	23.6	34.5	36.8	52.0
Skywork-OR1-Math (7B)	6.4	8.0	8.3	8.0	9.8	12.0	0.8	0.0	6.3	<u>7.0</u>	45.2	72.0	40.0	63.6	62.3	69.2	30.2	37.9	41.3	59.0
AceMath-RL (7B)	6.4	8.0	10.7	12.0	7.8	8.0	1.4	0.0	<u>6.6</u>	<u>7.0</u>	55.1	88.0	46.6	75.8	66.9	76.9	31.0	44.8	<u>46.9</u>	<u>70.0</u>
OpenMath-Nemo. (7B)	25.0	32.0	20.8	28.0	22.3	36.0	4.8	4.0	18.2	25.0	86.8	100.0	72.7	90.9	91.8	100.0	57.9	79.3	74.4	91.0
DS-R1-Distill (14B)	5.2	0.0	5.3	4.0	8.7	16.0	0.2	0.0	4.9	5.0	43.1	56.0	38.9	66.7	58.2	69.2	24.8	31.0	38.4	54.0
Light-R1-DS (14B)	6.2	4.0	7.5	8.0	10.9	12.0	0.2	0.0	<u>6.2</u>	<u>6.0</u>	56.6	84.0	45.5	75.8	66.5	76.9	28.7	37.9	<u>46.1</u>	<u>67.0</u>
OpenMath-Nemo. (14B)	28.7	32.0	26.1	40.0	26.8	40.0	4.2	4.0	21.4	29.0	88.3	100.0	75.2	100.0	94.5	100.0	60.2	86.2	76.6	96.0
Qwen3 (30B-A3B, Think)	35.6	40.0	24.1	28.0	18.1	24.0	2.7	4.0	20.1	24.0	87.8	92.0	84.7	97.0	91.3	100.0	61.9	65.5	79.7	87.0
DS-R1-Distill (32B)	6.5	0.0	5.4	4.0	10.6	12.0	0.7	0.0	5.8	4.0	45.2	52.0	41.8	63.6	60.2	69.2	26.0	37.9	40.4	54.0
QwQ (32B)	20.9	24.0	15.9	16.0	17.6	24.0	2.0	0.0	14.1	16.0	85.4	96.0	76.6	97.0	92.9	100.0	53.8	69.0	74.3	89.0
Light-R1-DS (32B)	16.8	28.0	12.0	12.0	13.4	16.0	4.4	16.0	11.6	<u>18.0</u>	70.1	96.0	64.1	93.9	80.4	92.3	39.8	51.7	60.7	82.0
OpenThinker2 (32B)	13.6	16.0	11.1	16.0	12.7	20.0	0.9	0.0	9.6	13.0	68.0	92.0	64.3	93.9	84.6	92.3	44.8	65.5	62.2	85.0
Skywork-OR1-Pre. (32B)	19.6	20.0	16.8	20.0	18.9	24.0	3.5	4.0	<u>14.7</u>	17.0	79.5	96.0	72.1	93.9	88.0	100.0	45.4	58.6	<u>68.3</u>	85.0
GLM-Z1-Air (32B)	18.0	16.0	12.0	8.0	16.0	16.0	2.5	4.0	12.1	11.0	76.0	96.0	69.3	78.8	89.4	92.3	41.8	48.3	65.6	76.0
OpenMath-Nemo. (32B)	22.5	36.0	22.5	32.0	22.5	28.0	3.5	4.0	17.8	25.0	68.0	96.0	62.5	90.9	90.4	100.0	48.7	72.4	63.5	<u>88.0</u>
Qwen3 (235B-A22B, Think)	36.5	48.0	43.5	48.0	28.5	32.0	4.0	8.0	28.1	34.0	91.0	100.0	90.2	97.0	94.2	100.0	78.4	89.7	87.5	96.0
DeepSeek R1	20.0	24.0	25.0	28.0	17.0	16.0	1.5	0.0	15.9	17.0	79.5	96.0	74.6	84.8	88.5	92.3	49.6	55.2	70.4	80.0
OpenAI o3-mini (high)	31.5	40.0	32.5	44.0	48.5	56.0	19.0	28.0	<u>32.9</u>	<u>42.0</u>	93.0	96.0	89.4	100.0	99.0	100.0	85.8	93.1	<u>90.5</u>	97.0
Gemini 2.5 Pro Exp 0325	65.0	76.0	78.0	80.0	53.5	56.0	25.0	40.0	<u>55.4</u>	<u>63.0</u>	90.5	96.0	93.2	93.9	100.0	100.0	84.1	86.2	90.8	<u>93.0</u>

Table 10: All models’ performance on OlymMATH (ZH). Models within each model size group are sorted by release time. The abbreviations “Alg.,” “Geo.,” “Num.,” and “Com.” represent the four categories in OlymMATH. Highest accuracy per model size is bolded. The second highest accuracy per model size is underlined. Models sampled only 8 times are marked in gray to indicate potential instability.

	Type	Example
Included	Real number	$2^{2017} + \arctan 2$
	Interval	$[\sqrt{33}, +\infty), (4, 5\pi]$
Excluded	Set Operations	$\{4, 5\} \cup \{1, 8\}$
	Variable	$\sqrt[3]{5a^2}, p^2 - pq$
	Complex number	$9 + 4i, \sqrt{-4}$
	Text	East, Alice

Table 11: *Included* and *excluded* formats of the answer.