RIMO: An Easy-to-Evaluate, Hard-to-Solve Olympiad Benchmark for Advanced Mathematical Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

As large language models reach high scores on benchmarks like GSM8K and MATH, researchers have started using Olympiad problems for new evaluations. However, grading these problems is difficult because of inconsistent answer formats and unreliable solutions. We present **RIMO**, a benchmark that keeps the challenging nature of Olympiad problems while ensuring clear and consistent evaluation. RIMO has two tracks: **RIMO-N**, which includes 335 problems redesigned to have single-integer answers for straightforward grading, and **RIMO-P**, which features 456 proof problems with expert-checked solutions and an automated grading system. Our results show that even the best LLMs struggle with RIMO, despite performing well on older benchmarks. RIMO reveals a significant gap in current models' reasoning abilities and offers a precise tool for future research.

1 Introduction

Large language models (LLMs) have shown striking progress in mathematical reasoning. However, early benchmarks like GSM8K [1] and MATH [2] are now saturated, with frontier systems surpassing 90% accuracy. Consequently, to probe the upper limits of machine reasoning, the research community has turned to the challenges posed by the International Mathematical Olympiad (IMO).

Recent efforts use Olympiad material, but practical constraints blur the evaluation signal. The **AIMO** [3] competition, for instance, is a dynamic contest with a hidden test set, limiting reproducibility. Static benchmarks like OLYMMATH [4] and OMNI-MATH [5] use heterogeneous answer formats (e.g., fractions, proofs) that require noisy evaluation via LLM-based judges, introducing bias and masking true capabilities.

This paper introduces RIMO (Remade International Mathematical Olympiad), a benchmark designed to preserve peak Olympiad difficulty while eliminating this evaluation noise. RIMO is a two-track benchmark built from IMO material from 1959 to 2023:

- **RIMO-N** consists of 335 problems, remade to yield a single, unique integer answer, allowing for deterministic, $\mathcal{O}(1)$ string-match grading.
- **RIMO-P** contains 456 proof problems, decomposed into a sequence of sub-problems to evaluate the step-by-step reasoning process.

We tested ten leading LLMs, such as GPT-4o and Gemini 2.5 Flash. While these models perform well on previous benchmarks, their scores fall sharply on RIMO. This shows there is still a large gap between what current LLMs can do and true Olympiad-level reasoning. By combining a clear integer-answer track with a carefully graded proof track, RIMO gives researchers a precise way to measure and address this reasoning gap.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

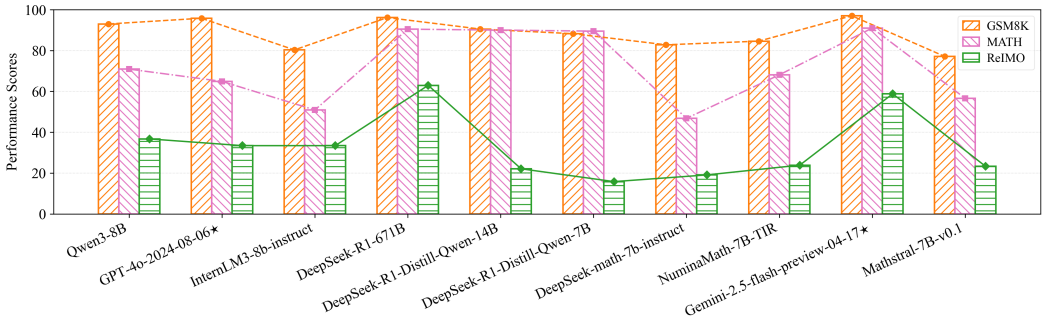


Figure 1: Comparisons among different models on GSM8K, MATH, and RIMO-N, where the models are ranked based on their performance on MATH, and those marked with "⋆" are closed-source models.

2 Benchmark Construction

RIMO is assembled entirely from International Mathematical Olympiad (IMO) material published between 1959 and 2023. For every year we gathered both the public contest paper and the confidential shortlist booklet, digitised the statements, and collated all available solutions. Each problem then passes through a verification-selection pipeline (Fig. 2): solutions are reconciled across multiple sources, the statement is either retained in its original form or carefully rewritten, and the finalised item is deposited in one of two tracks, **RIMO-N** or **RIMO-P**.

2.1 RIMO-N: single-integer problems

The RIMO-N track comprises 335 problems, 236 drawn from shortlist booklets and 99 from contest papers, that have been remade so each admits a single, unique integer answer. Remaking is never a cosmetic tweak to the last line: intermediate hypotheses are tightened when ambiguity appears, variables are renamed for coherence, and objectives are reframed, yet the logical core and difficulty of the source problem remain intact. A concurrency proof, for instance, may become "how many common points do the circumcircles have", while a classification of integer triples can be recast as "compute the value of $a + b + c$ over every such triple." Figure 3 shows two representative transformations. Content-wise the set stays faithful to traditional IMO proportions, covering algebra (96 items), geometry (95), number theory (86), and combinatorics (58).

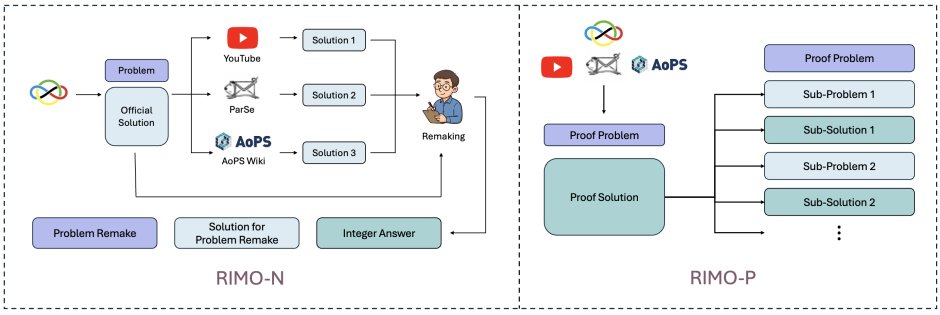


Figure 2: End-to-end construction pipeline of RIMO-N and RIMO-P.

Each shortlist problem retains the jury's official integer. A contest problem, lacking an authorised key, is accepted only when at least two of three independent community sources-AoPS Wiki, YouTube blackboard expositions, and ParSe transcripts—return exactly the same answer; any disagreement triggers manual adjudication and usually leads to exclusion. With this guarantee in place, grading collapses to a constant-time string comparison, freeing RIMO-N from symbolic post-processing or learned judges.

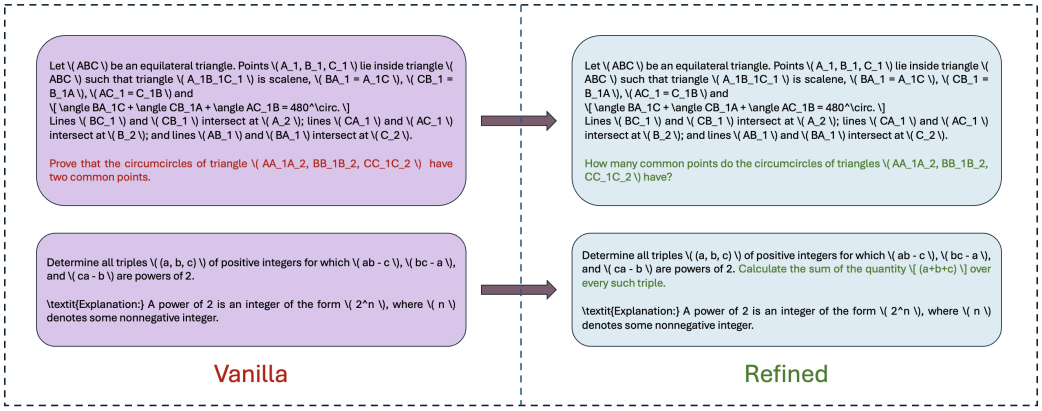


Figure 3: **Vanilla vs. Refined questions.** Grey text is copied verbatim from the original statement; the coloured line is rewritten so the answer becomes a unique integer.

2.2 RIMO-P: Decomposed Proof Problems

Where RIMO-N targets deterministic answer checking, RIMO-P is designed to measure a model's capacity for the **process** of full deductive reasoning. The 456 problems in this track are decomposed into a sequence of guided sub-problems. To create this structure, we use expert-verified proofs from official IMO shortlists and community sources (Figure 2). The word count of a reference proof determines its decomposition into one to four sub-problems, with longer solutions typically yielding a three or four-step logical pathway. This scaffolded design allows for a granular evaluation of a model's ability to solve intermediate lemmas, while the final sub-problem in every sequence preserves the original problem's main goal, offering deeper insight into its deductive capabilities.

3 Evaluation and Insights

A benchmark's utility depends on reliable evaluation. Unlike benchmarks that suffer from "evaluation noise" due to ambiguous answer formats (Figure 8), RIMO is designed for robustness. **RIMO-N**'s single-integer format allows for deterministic string-match grading, while **RIMO-P**'s decomposed structure enables a clear, step-by-step assessment of logical reasoning. We evaluated ten models using greedy decode ($T = 0$) to ensure reproducibility.

3.1 Evaluation on RIMO

On **RIMO-N**, all models exhibit a drastic performance drop compared to their scores on MATH and GSM8K (Table 4a). This exposes a significant gap between solving standard competition math and true Olympiad-level problems.

For **RIMO-P**, we use a sequential protocol judged by **deepseek-r1 (the top RIMO-N model)**, where a proof is graded on the number of consecutive sub-problems solved correctly. The final performance score, P, is the average proportion of completed steps, calculated as:

$$P = \frac{1}{N} \sum_{i=1}^{N} \frac{S_i}{X_i} \tag{1}$$

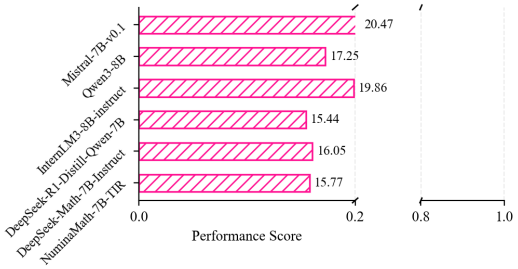
where N is the total number of problems (456), S_i is the number of consecutively correct sub-solutions for problem i, and X_i is its total number of sub-problems. Performance on this track is very low (Figure 4b), and strong RIMO-N scores do not guarantee success. This indicates that answer-finding and rigorous proof-writing are distinct capabilities that current models struggle with.

3.2 Key Insights from RIMO Analysis

Our detailed analysis of the RIMO results provides several crucial insights into the current state of mathematical LLMs. First, progress is not simply a function of size or recency; massive scale appears

(a) Pass@1 accuracy (%) of models on three benchmarks.

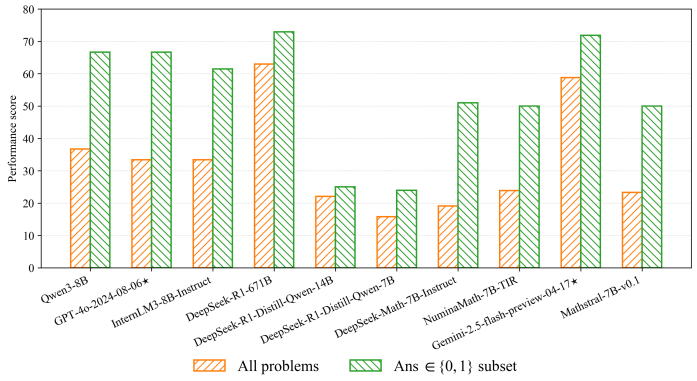
Model	GSM8K	MATH	RIMO
Qwen3-8B	93.00	70.90	36.72
GPT-4o-2024-08-06⋆	95.80	64.88	33.43
InternLM3-8B-instruct	80.30	50.90	33.43
DeepSeek-R1-671B	96.13	90.45	62.96
DS-R1-Distill-Q-14B	90.50	90.20	22.09
DS-R1-Distill-Q-7B	88.24	89.49	15.82
DeepSeek-math-7B	82.80	46.80	19.10
NuminaMath-7B-TIR	84.60	68.10	23.88
Gemini-2.5-flash⋆	97.04	91.31	58.81
Mathstral-7B-v0.1	77.10	56.60	23.28



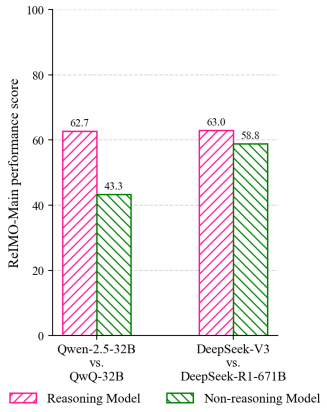
(b) The performance scores on RIMO-P.

Figure 4: Overall performance comparison on RIMO-N (left) and RIMO-P (right).

to be a prerequisite for top performance, but **targeted data and architectural choices** are dominant factors otherwise. Second, models explicitly optimized for reasoning consistently outperform their vanilla counterparts of the same size, confirming that **specialized training yields tangible gains** (Figure 5b) at the Olympiad level. Finally, the nature of the problem itself is critical. All models perform substantially better on questions with binary (0/1) answers (Figure 5a), revealing that a significant portion of RIMO's challenge comes from forcing models to **locate a precise integer in a large numerical space**, rather than simply deciding a true/false claim. These findings suggest that future breakthroughs will likely stem from a combination of scale, specialized reasoning training, and improved numerical search capabilities.



(a) Pass@1 on full RIMO-N vs. binary subset.



(b) RIMO-N accuracy: reasoning-optimised vs. vanilla models.

Figure 5: Comparative RIMO-N analysis: (a) impact of binary outputs, (b) effect of reasoning optimization.

4 Discussion and Future Work

Our results on RIMO clearly show the reasoning gap between current LLMs and Olympiad-level mathematics. The different outcomes on RIMO-N and RIMO-P indicate that **answer-finding** and **rigorous proof-writing** are distinct skills. Our evaluation highlights this gap, though limited access to GPUs and the high cost of proprietary model APIs restricted us from testing more state-of-the-art models. Even so, RIMO's noise-free framework remains a dependable way to track real progress as these models develop.

Future work will build on this foundation. First, we will translate RIMO-P into a formal language like **LEAN** to create a benchmark for machine-verifiable proofs—the highest standard of correctness. We will also expand our leaderboard as more models become accessible. Finally, RIMO-P's granular structure enables detailed error analysis to pinpoint specific model weaknesses, guiding the development of more capable and reliable AI systems.

References

[1] 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. Training verifiers to solve math word problems, 2021.

[2] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021.

[3] Simon Frieder, Sam Bealing, Arsenii Nikolaiev, Geoff C. Smith, Kevin Buzzard, Timothy Gowers, Peter J. Liu, Po-Shen Loh, Lester Mackey, Leonardo de Moura, Dan Roberts, D. Sculley, Terence Tao, David Balduzzi, Simon Coyle, Alex Gerko, Ryan Holbrook, Addison Howard, and XTX Markets.

[4] Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models, 2025.

[5] 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. Omni-math: A universal olympiad level mathematic benchmark for large language models, 2024.

[6] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022.

[7] Daman Arora, Himanshu Gaurav Singh, and Mausam. Have llms advanced enough? a challenging problem solving benchmark for large language models, 2023.

[8] Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data, 2024.

[9] Trieu Trinh, Yuhuai Wu, Quoc Le, He He, and Lng Thng. Solving olympiad geometry without human demonstrations. *Nature*, 625:476–482, 01 2024.

[10] Yujun Mao, Yoon Kim, and Yilun Zhou. Champ: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities, 2024.

[11] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

[12] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai, 2025.

A Details on the related work

A.1 Mathematics benchmarks

Early efforts to quantify the mathematical competence of large language models primarily focused on material from grade school or high school. GSM8K evaluates multi-step word problems that require nothing beyond basic arithmetic and algebra; frontier models now exceed 95 % accuracy, making the dataset largely saturated [1]. The MATH dataset extended the scope to 12,500 high-school competition questions—many drawn from AMC and AIME examinations—yet recent proprietary systems (e.g.,

OpenAI o1-mini) already solve more than 90 % of its items, limiting its diagnostic power [2]. To restore headroom, several groups proposed harder, mixed-level corpora. OCWCourses gathers 272 undergraduate STEM questions from MIT OpenCourseWare [6]; JEEBench selects 515 IIT-JEE Advanced problems spanning mathematics, physics, and chemistry [7], and MathOdyssey combines university calculus with Olympiad-style reasoning to expose failure modes once routine patterns are removed [8]. Although these datasets are demonstrably more challenging than GSM8K and MATH, most still feature heterogeneous answer formats—such as fractions, radicals, and intervals—that complicate exact-match scoring and sometimes necessitate model-based evaluation.

A.2 Olympiad-level benchmarks

A parallel line of work turns directly to International-level competitions, whose problems demand long deductive chains and creative insights. AlphaGeometry restricts itself to synthetic and real Olympiad geometry [9], while CHAMP supplies 270 high-school contest problems annotated with key concepts and hints [10]. OlympiadBench broadens the domain to mathematics and physics, presenting 8,476 bilingual items but relying on GPT-4V to adjudicate answers when closed-form checking fails [11]. OlympicArena mixes mathematics with other cognitively demanding puzzles and again adopts a model-based evaluation pipeline [12].

Two recent datasets focus exclusively on text-only Olympiad mathematics. OlymMATH introduces 200 bilingual IMO-style problems divided into AIME-level easy and genuine Olympiad complex subsets; answers are numeric yet still include expressions such as $\sqrt{4+\sqrt{5}}$ or open intervals, so symbolic equivalence logic is required for grading [4]. Omni-MATH scales the idea to 4428 problems, covering 33 sub-domains and 10 difficulty tiers, but must employ GPT-4o and an auxiliary Omni-Judge model to handle diverse output forms [5]. Although both benchmarks significantly increase difficulty, the dependence on expression normalization or learned judges introduces evaluation noise and potential bias.

A.3 Positioning RIMO

Our work, RIMO, inherits the Olympiad focus of OlymMATH and Omni-MATH but targets their principal limitation: ambiguous grading. By remaking 335 International Mathematical Olympiad problems—spanning from 1959 to 2023—so that each admits a single, unique integer answer, RIMO restores deterministic, rule-based evaluation while maintaining genuine Olympiad difficulty. Multi-source cross-checking (official shortlist solutions, AoPS-Wiki write-ups, YouTube expositions, and ParSe transcripts) further ensures the reliability of ground truth. Consequently, RIMO provides a clean, high-resolution yardstick for measuring the next generation of reasoning-centric large language models (LLMs).

B Examples of the Problems in RIMO-N and RIMO-P

Problem: For every integer $n \geq 1$ consider the $n \times n$ table with entry $\left\lfloor \frac{ij}{n+1} \right\rfloor$ at the intersection of row i and column j, for every $i = 1, \ldots, n$ and $j = 1, \ldots, n$. Determine the smallest integers $n \geq 1$ for which the sum of the n^2 entries in the table is equal to $\frac{1}{4}n^2(n-1)$ and n is not a prime.

Solution: First, observe that every pair x, y of real numbers for which the sum $x + y$ is integer satisfies

$$|x| + |y| \geq x + y - 1. \tag{1}$$

The inequality is strict if x and y are integers, and it holds with equality otherwise. We estimate the sum S as follows.

$$2S = \sum_{1 \leq i,j \leq n} \left(\left\lfloor \frac{ij}{n+1} \right\rfloor + \left\lfloor \frac{ij}{n+1} \right\rfloor \right) = \sum_{1 \leq i,j \leq n} \left(\left\lfloor \frac{ij}{n+1} \right\rfloor + \left\lfloor \frac{(n+1-i)j}{n+1} \right\rfloor \right)$$

$$\geq \sum_{1 \leq i,j \leq n} (j-1) = \frac{(n-1)n^2}{2}.$$

The inequality in the last line follows from (1) by setting $x = \frac{ij}{n+1}$ and $y = \frac{(n+1-i)j}{n+1}$, so that $x + y = j$ is integral.
Now $S = \frac{1}{4}n^2(n-1)$ if and only if the inequality in the last line holds with equality, which means that none of the values $\frac{ij}{n+1}$ with $1 \leq i, j \leq n$ may be integral.
Hence, if $n+1$ is composite with factorisation $n+1 = ab$ for $2 \leq a, b \leq n$, one gets a strict inequality for $i = a$ and $j = b$. If $n+1$ is a prime, then $\frac{ij}{n+1}$ is never integral and $S = \frac{1}{4}n^2(n-1)$. Since n is not a prime, the answer is 4.

Answer: 4

Type: algebra

Figure 6: Example of the Problems in RIMO-N.

Problem ID: 2023a3

Problem: Let $x_1, x_2, \ldots, x_{2023}$ be distinct real positive numbers such that

$$a_n = \sqrt{(x_1 + x_2 + \cdots + x_n)\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}$$

is an integer for every $n = 1, 2, \ldots, 2023$. Prove that $a_{2023} \geqslant 3034$.

Parts: 2

Sub-Problem 1: Let the sequence a_n be defined as in the problem. Show that if $a_{n+1} - a_n = 1$ for some $n \geq 1$, then $a_{n+2} - a_{n+1} \geq 2$.

Sub-Solution 1: We have the identity $a_{k+1}^2 = (\sum_{i=1}^{k+1} x_i)(\sum_{i=1}^{k+1} \frac{1}{x_i}) = a_k^2 + 1 + (\sum_{i=1}^{k} x_i)\frac{1}{x_{k+1}} + x_{k+1}(\sum_{i=1}^{k} \frac{1}{x_i})$. By AM-GM inequality, $(\sum_{i=1}^{k} x_i)\frac{1}{x_{k+1}} + x_{k+1}(\sum_{i=1}^{k} \frac{1}{x_i}) \geq 2\sqrt{(\sum x_i)(\sum \frac{1}{x_i})} = 2a_k$. Thus, $a_{k+1}^2 \geq a_k^2 + 2a_k + 1 = (a_k + 1)^2$, which implies $a_{k+1} \geq a_k + 1$. The condition $a_{k+1} - a_k = 1$ is equivalent to the equality case of the AM-GM, which is $(\sum_{i=1}^{k} x_i)\frac{1}{x_{k+1}} = x_{k+1}(\sum_{i=1}^{k} \frac{1}{x_i})$. Assume for contradiction that $a_{n+1} - a_n = 1$ and $a_{n+2} - a_{n+1} = 1$. This means the equality condition holds for both $k = n$ and $k = n + 1$. This leads to the equations $\frac{1}{x_{n+1}} \sum_{i=1}^{n} x_i = x_{n+1} \sum_{i=1}^{n} \frac{1}{x_i}$ and $\frac{1}{x_{n+2}} \sum_{i=1}^{n+1} x_i = x_{n+2} \sum_{i=1}^{n+1} \frac{1}{x_i}$. Manipulating these two equations leads to $\sum_{i=1}^{n} \frac{1}{x_i} = -\frac{1}{x_{n+1}}$, a contradiction since all x_i are positive. Thus, two consecutive increments of 1 are not possible. Since $a_{n+2} - a_{n+1}$ must be an integer and at least 1, it follows that if $a_{n+1} - a_n = 1$, then $a_{n+2} - a_{n+1} \geq 2$.

Sub-Problem 2: Let $x_1, x_2, \ldots, x_{2023}$ be distinct real positive numbers such that

$$a_n = \sqrt{(x_1 + x_2 + \cdots + x_n)\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right)}$$

is an integer for every $n = 1, 2, \ldots, 2023$. Prove that $a_{2023} \geqslant 3034$.

Sub-Solution 2: We know $a_1 = 1$. Let $\Delta_n = a_{n+1} - a_n$ for $n = 1, \ldots, 2022$. Since the x_i are distinct and positive, a_n is a strictly increasing sequence of integers, so $\Delta_n \geq 1$. From the previous sub-problem, we know that if $\Delta_n = 1$, then $\Delta_{n+1} \geq 2$. We have $a_{2023} = a_1 + \sum_{n=1}^{2022} \Delta_n = 1 + \sum_{n=1}^{2022} \Delta_n$. We can group the 2022 differences into 1011 pairs: $(\Delta_{2k-1} + \Delta_{2k})$ for $k = 1, \ldots, 1011$. For each pair, if $\Delta_{2k-1} = 1$, then $\Delta_{2k} \geq 2$, so their sum is at least 3. If $\Delta_{2k-1} \geq 2$, then since $\Delta_{2k} \geq 1$, their sum is also at least 3. In all cases, $\Delta_{2k-1} + \Delta_{2k} \geq 3$. Thus, the total sum of the differences is $\sum_{n=1}^{2022} \Delta_n = \sum_{k=1}^{1011} (\Delta_{2k-1} + \Delta_{2k}) \geq \sum_{k=1}^{1011} 3 = 1011 \times 3 = 3033$. Therefore, $a_{2023} = 1 + \sum_{n=1}^{2022} \Delta_n \geq 1 + 3033 = 3034$.

Figure 7: Example of the Problems in RIMO-P.

C An example of how AI judger fails

Domain: Mathematics → Algebra → Linear Algebra → Matrices

Difficulty: 9.0 (Given by Omni-Math)

Source: Putnam

Problem: Let n be a positive integer. For i and j in $\{1, 2, \ldots, n\}$, let $s(i, j)$ be the number of pairs (a, b) of nonnegative integers satisfying $ai + bj = n$. Let S be the n-by-n matrix whose (i, j) entry is $s(i, j)$. For example, when $n = 5$, we have $S = \begin{bmatrix} 6 & 3 & 2 & 2 & 2 \\ 3 & 0 & 1 & 0 & 1 \\ 2 & 1 & 0 & 0 & 1 \\ 2 & 0 & 0 & 0 & 1 \\ 2 & 1 & 1 & 1 & 2 \end{bmatrix}$. Compute the determinant of S.

Answer: $(-1)^{\lceil n/2 \rceil - 1} 2 \lceil \frac{n}{2} \rceil$

Predict:

$$f(n) = \begin{cases} (-1)^{\frac{n}{2}-1} \cdot n & \text{if } n \text{ is even} \\ (-1)^{\frac{n-1}{2}} \cdot (n+1) & \text{if } n \text{ is odd} \end{cases}$$

Equivalence Judgement: FALSE

Justification: The reference answer gives the determinant as:

$$(-1)^{\lceil n/2 \rceil - 1} \cdot 2 \cdot \lceil n/2 \rceil$$

This can be broken down according to the parity of n:

- **If n is even**, then $\lceil n/2 \rceil = n/2$, so $(-1)^{n/2-1} \cdot 2 \cdot \frac{n}{2} = (-1)^{n/2-1} n$.
- **If n is odd**, then $\lceil n/2 \rceil = (n+1)/2$, so $(-1)^{(n+1)/2-1} \cdot 2 \cdot \frac{n+1}{2} = (-1)^{(n-1)/2} (n+1)$.

Hence the expression simplifies to the piecewise form

$$f(n) = \begin{cases} (-1)^{\frac{n}{2}-1} n & \text{if } n \text{ is even,} \\ (-1)^{\frac{n-1}{2}} (n+1) & \text{if } n \text{ is odd.} \end{cases}$$

This is *identical* to the student's boxed expression. However, according to the rubric, the student's answer must match the reference answer's boxed expression *exactly*, namely $(-1)^{\lceil n/2 \rceil - 1} \cdot 2 \cdot \lceil n/2 \rceil$. The student instead rewrote it in a piecewise form depending on the parity of n. Although mathematically equivalent, **this does not follow the instruction requiring exact correspondence to the reference answer's boxed content**.

Figure 8: An example of evaluation noise from an LLM-based judge (GPT-4o) on the Omni-MATH benchmark. The judge incorrectly marked the prediction of a model as "FALSE" despite it being mathematically identical to the reference answer, penalizing a stylistic difference in format.

D Details on key insights

Are Newer Models Automatically Stronger? Chronology tells an even messier story (Fig. 9). DeepSeek-Math-7B (Feb 2024) beats the August-2024 Mathstral-7B despite being half a year older. Conversely, Gemini 2.5-flash (Apr 2025) nearly matches the much larger DeepSeek-R1-671B released two months earlier. Across the ten points the Spearman rank correlation between publication date and accuracy is only 0.21. Incremental architectural tweaks or enlarged instruction corpora therefore do not guarantee progress on Olympiad mathematics; breakthroughs seem to coincide with either massive scale (DeepSeek-671B) or targeted domain pre-training (Gemini 2.5-flash, Qwen3-8B).

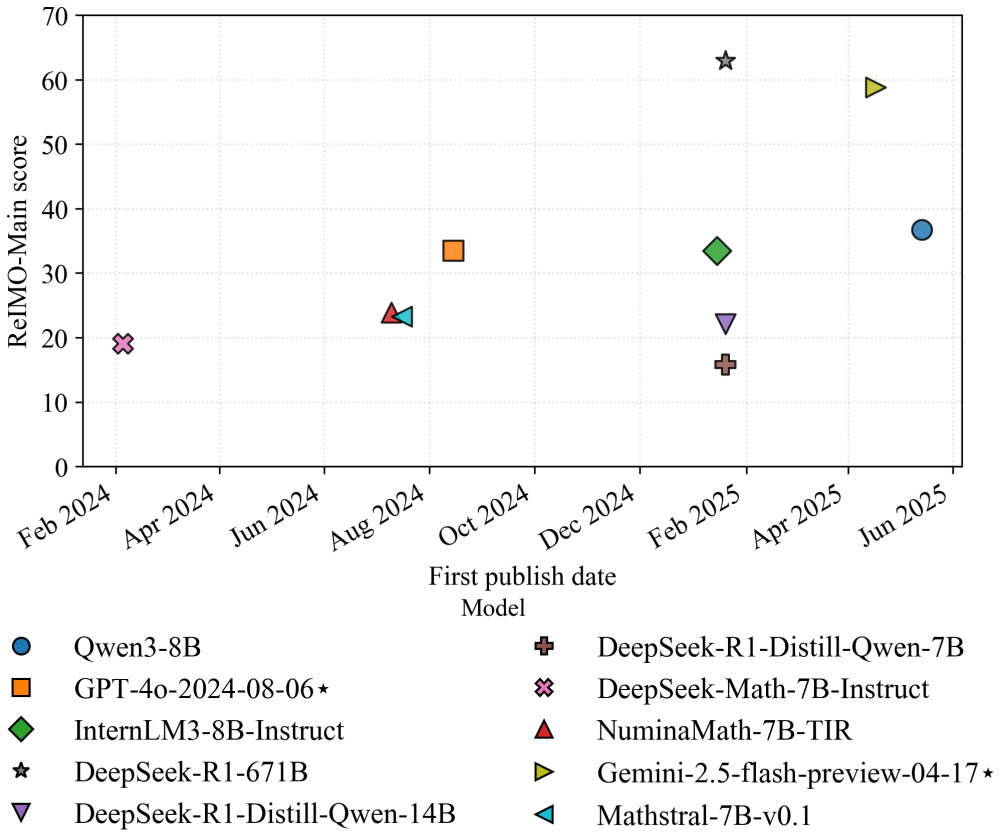


Figure 9: RIMO-N accuracy versus first-release date.

While massive scale appears necessary for state-of-the-art scores, it does not guarantee top performance. Below a certain threshold, factors like training data and objectives appear to dominate raw parameter count.

Does Bigger Still Mean Better? Figure 10 places all ten models on a log-parameter axis. The only model whose size is published far above 100 B—DeepSeek-R1-671B—indeed tops the chart at 63%. For the two proprietary systems whose parameter counts remain undisclosed (GPT-4o and Gemini 2.5-flash) we plot them alongside DeepSeek-R1 in the "very-large" regime. Gemini almost matches DeepSeek's score (59%), whereas GPT-4o reaches only 33%, underscoring that whatever its scale, sheer width does not guarantee Olympiad prowess. Within the disclosed 7–14B cluster the pattern is even clearer: Qwen3-8B (37%) and InternLM3-8B (33%) both outperform the larger 14 B distilled checkpoint (22%) and several maths-specialised 7B models (<24%). Taken together, the scatter suggests a threshold effect—massive scale is *necessary* for state-of-the-art scores, but below that threshold training data and objective dominate; once in the very-large regime, architectural choices and domain pre-training still separate winners from also-rans.

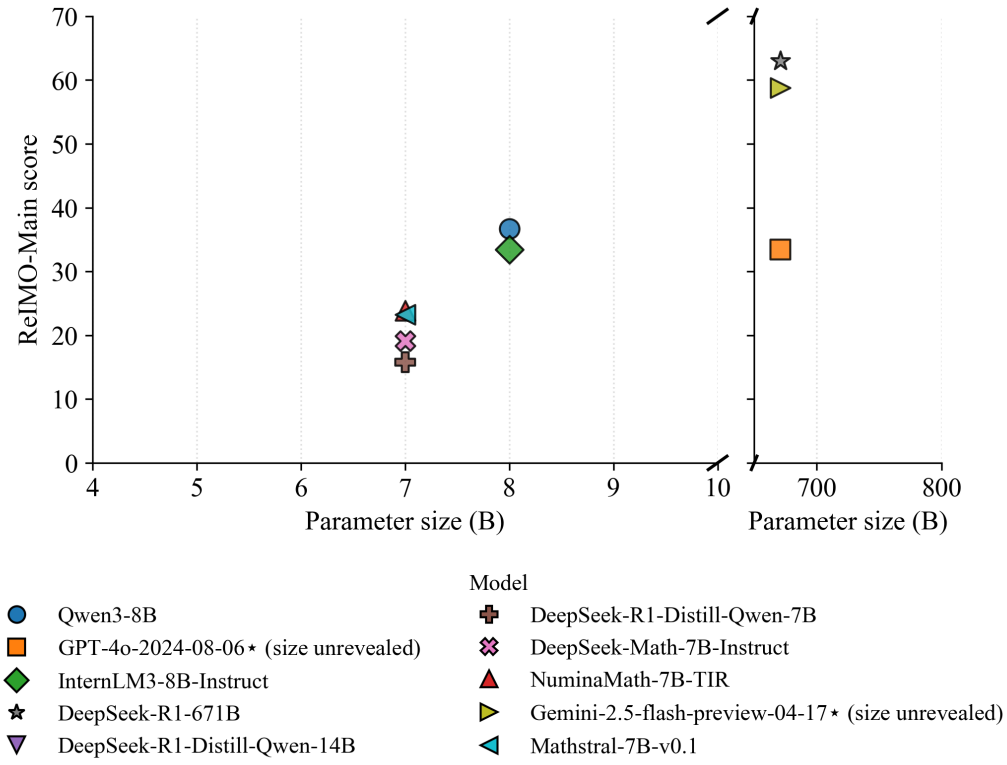


Figure 10: RIMO-N accuracy versus parameter size.

Breakthroughs appear to coincide with either massive scale increases or targeted domain pre-training, not incremental release updates.

Does Answer Sparsity Change the Game? Binary-valued items—in Olympiad terms, many of these are *true-or-false* statements whose proof reduces to deciding whether the claim is correct—substantially relax the search space for an LLM. RIMO-N contains 96 such problems whose ground-truth integer happens to be 0 or 1. Roughly two-thirds of them are genuine T/F formulations (e.g. "prove that the two circumcircles have no common point"), while the remainder still ask for a numeric extremum that just evaluates to 0 or 1. Figure 11 plots accuracy on this "binary subset" alongside full-set accuracy. Figure 11 illustrates the effect. Scores jump by 8 to 30 percentage pointsacross in every baseline: DeepSeek-R1-671B climbs from 63% to 73%, Qwen3-8B from 37% to 67%, and the weakest system, DeepSeek-R1-Distill-Qwen-7B, rises from 16% to 24%. While random guessing yields 50% on a strict T/F task, the persistent margin above chance shows that models exploit more than luck, yet the consistent gap confirms that having only two admissible outputs removes a significant portion of RIMO's challenge. In other words, part of the benchmark's hardness comes from forcing models to locate the exact integer on a larger numerical spectrum—not merely to affirm or deny a statement.

A significant portion of the challenge of RIMO challenge comes from forcing models to locate an integer in a large numerical space. Restricting the answers to a binary choice substantially inflates accuracy across all models.

Reasoning Model vs. Non-reasoning Model: Recent releases such as QwQ-32B and DeepSeek-R1 adopt explicit reasoning objectives (self-refinement, chain-of-thought distillation, or specialized reward modeling) on top of a backbone shipped in a "plain" form. Figure 12 contrasts each reasoning model with its non-reasoning sibling at an identical or near-identical scale. On RIMO-N the reasoning variants consistently win: QwQ-32B outperforms Qwen-2.5-32B by **19.4 percentage points** (62.7 vs. 43.3) and DeepSeek-R1-671B edges out the newly released DeepSeek-V3 by **4.2 percentage points**

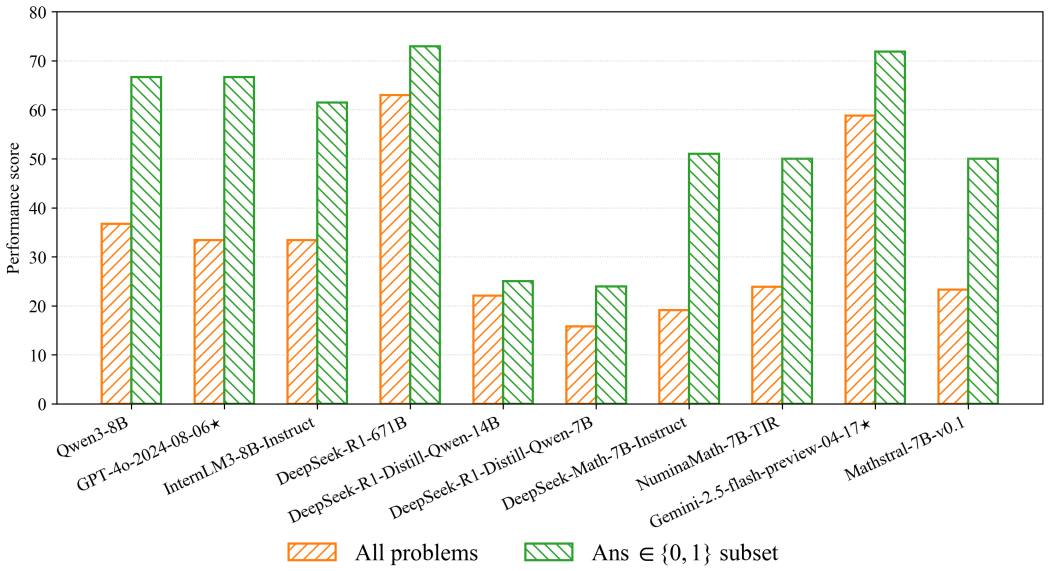


Figure 11: Pass@1 on the full RIMO-N (dark bars) versus on the 96-problem subset whose answers lie in $\{0, 1\}$ (light bars). The systematic boost highlights how much easier binary outputs are for current LLMs.

(63.0 vs. 58.8). The margin is huge when the base model is instruction-oriented but not maths-centric (Qwen-2.5). These results indicate that explicit reasoning optimization yields tangible gains even at the Olympiad level, over and above what scale or generic instruction tuning alone can offer.

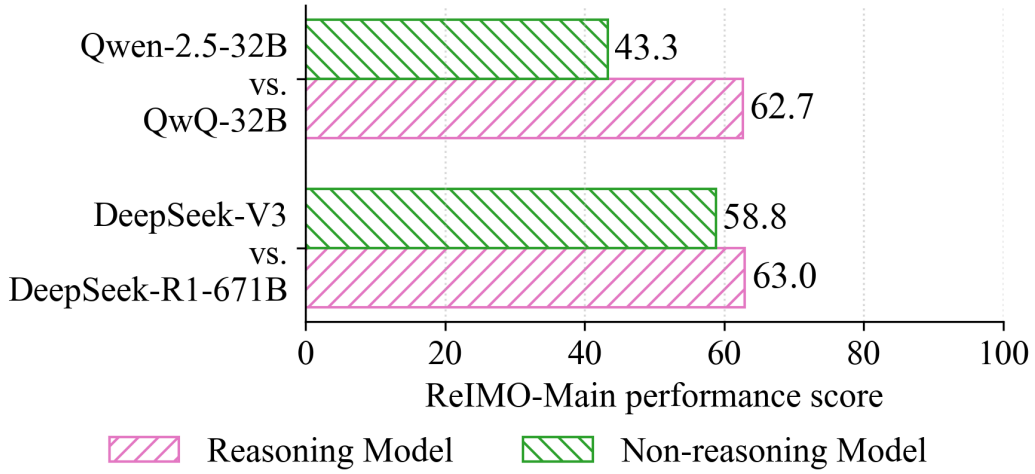


Figure 12: RIMO-N accuracy: reasoning-optimised models (dark bars) vs. size-matched vanilla counterparts (light bars).

Explicit optimization for reasoning provides tangible gains at the Olympiad level, offering performance improvements over and above what scale or generic instruction tuning alone can provide.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions in advanced math benchmarks.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: Our paper explicitly discusses that its evaluation was limited by available GPUs and the high cost of proprietary model APIs, which constrained the range of state-of-the-art systems tested.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper introduces an empirical benchmark and evaluates model performance, rather than presenting theoretical results that would require mathematical proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper discloses the details needed to reproduce its main experimental findings, including the benchmark's construction methodology, a zero-temperature greedy strategy, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provides open access to the RIMO dataset.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper assess the pre-trained models as provided with their default parameters with zero-temperature.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our paper reports performance using deterministic point estimates of accuracy on the benchmark and does not provide error bars or an analysis of statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

Justification: The experiments consist of evaluating pre-trained models, many of which are accessed via APIs, so no specific local computational resources are required for reproduction.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Our research fully conforms with the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our paper shows its contribution as a positive impact for the AI research community. And no potential negative societal impact is shown in our research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper releases a benchmark dataset of public mathematical problems, not a high-risk model or dataset that would necessitate the discussion of specific misuse safeguards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Our paper properly credits the creators of the data and models used by citing their original sources and distinguishing between open and proprietary systems.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, the RIMO is well-documented within the paper itself, and this documentation is provided alongside the dataset and code.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper evaluates AI models using existing public data and does not conduct any new crowdsourcing experiments or research involving human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our research evaluates AI systems on a mathematical benchmark and does not involve human subjects, thus requiring no IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Our paper's core methodology describes the usage of LLMs in two key ways: as the subjects being evaluated on the benchmark, and as an essential component for automatically grading the proof-based track.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.