AMO-Bench: Large Language Models Still Struggle in High School Math Competitions

Anonymous ACL submission

Abstract

We present AMO-Bench, an **A**dvanced **M**athematical reasoning benchmark with **O**lympiad level or even higher difficulty, comprising 50 human-crafted problems. Existing benchmarks have widely leveraged high school math competitions for evaluating mathematical reasoning capabilities of large language models (LLMs). However, many existing math competitions are becoming less effective for assessing top-tier LLMs due to performance saturation (e.g., AIME24/25). To address this, AMO-Bench introduces more rigorous challenges by ensuring all 50 problems are (1) cross-validated by experts to meet at least the International Mathematical Olympiad (IMO) difficulty standards, and (2) entirely original problems to prevent potential performance leakages from data memorization. Experimental results across 36 LLMs on AMO-Bench highlights three key findings: (1) high-level mathematical reasoning remains challenging for current LLMs, with even the best-performing model achieving only 63.1% accuracy and most LLMs scoring below 50%; (2) scaling test-time compute remains a highly effective strategy for substantially improving reasoning performances, and (3) open-source models are progressively narrowing the performance gap with proprietary models. Additionally, we conduct further analysis about reasoning efficiency, volatility, and cross-lingual robustness, providing deeper insights behind the reasoning performances.

1 Introduction

Recent advances in LLMs have demonstrated significant improvements in reasoning capabilities (OpenAI, 2024; Gemini Team, 2025; OpenAI, 2025; Anthropic, 2025; xAI, 2025; Yang et al., 2025; Guo et al., 2025; DeepSeek-AI, 2025; Meituan LongCat Team, 2025a; GLM-4.5 Team, 2025; ByteDance Seed, 2025; Tencent Hunyuan Team, 2025; Kimi Team, 2025; Meituan LongCat Team, 2025b). To track this rapid progress, mathematical problem solving has become a critical metric for evaluation, as it inherently demands complex and multi-step reasoning processes to arrive at correct answers. As a result, many current benchmarks utilize problems from high school mathematics competitions (e.g., HMMT and AIME) to assess the reasoning abilities of LLMs (Balunović et al., 2025; He et al., 2024; Gao et al., 2024; Fang et al., 2025). Recent results indicate that state-of-the-art models are achieving remarkable performances on these benchmarks, with some even surpassing 90% accuracy on competitions like AIME24/25.

However, these impressive results also expose an emerging challenge: many existing mathematics benchmarks are approaching performance saturation and are becoming less effective for assessing further advancements in reasoning capabilities. On the one hand, as LLMs gradually approach or even surpass human-level capabilities in mathematics, some math competitions are becoming less challenging for top-tier models (OpenAI, 2025; DeepSeek-AI, 2025; Yang et al., 2025; Meituan LongCat Team, 2025b). On the other hand, most current benchmarks are derived from previous competitions, raising concerns about potential data memorization and performance leakage (Sun et al., 2025; Balunović et al., 2025). While recent efforts have incorporated problems from more difficult and newly held contests such as the International Mathematical Olympiad (IMO), these questions tend to be proof-based and require manual verification by experts (Balunović et al., 2025; Petrov et al., 2025). This reliance on expert review hinders the implementation of automated scoring processes, leading to inefficiency and inconsistency in large-scale evaluations and result reproductions.

To address these limitations, we present AMO-Bench, an advanced mathematical reasoning benchmark consisting of 50 novel and extremely challenging problems. The core features of AMO-

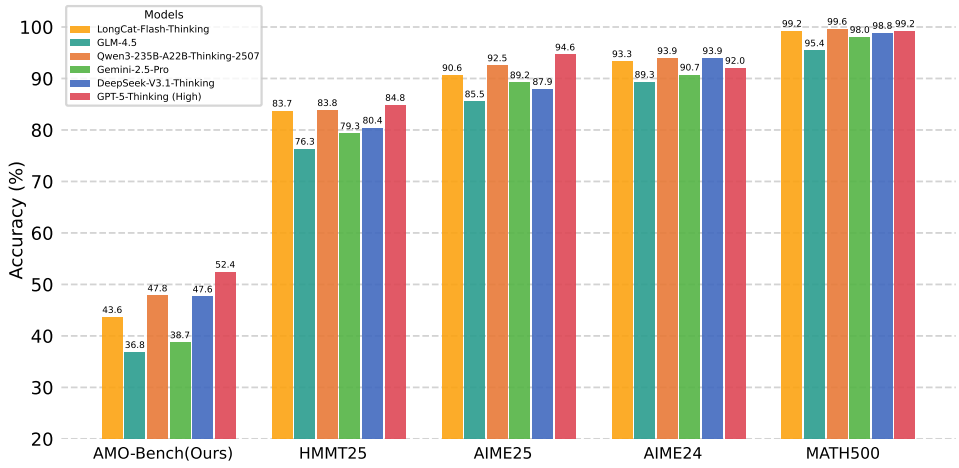


Figure 1: Performance of LLMs on AMO-Bench as well as existing competition-level math benchmarks. Except for the results on AMO-Bench, all other results are sourced from Meituan LongCat Team (2025b).

Bench are as follows: **(1) Original problems.** All problems in AMO-Bench are newly crafted by human experts to prevent performance leaks from existing resources, with secondary verification to ensure originality. **(2) Guaranteed difficulty.** Each problem has undergone rigorous cross-validation by multiple experts to ensure it meets at least the difficulty standards of IMO, and an LLM-based difficulty filtering stage excludes questions that do not challenge current reasoning models. **(3) Final-answer based grading.** Each problem in AMO-Bench requires a final answer, enabling efficient automatic grading using parser-based or LLM-based method according to the answer type, balancing grading cost and generalizability. **(4) Human-annotated reasoning paths.** Each problem includes a detailed reasoning path written by human experts, enhancing solution transparency and supporting further explorations on AMO-Bench, such as prompt engineering and error analysis.

Experimental results across various LLMs demonstrate that contemporary LLMs still struggle with the significant challenges presented by AMO-Bench. Among 36 evaluated models, the state-of-the-art accuracy on AMO-Bench is only 63.1%, achieved by Gemini-3-Pro, with most models scoring below 50%. Figure 1 illustrates the performance of several models on AMO-Bench as well as the comparison with other mathematical benchmarks. Despite the limited performances of current LLMs, our analysis also reveals considerable potential for further improvements. We show that the model performances exhibit a near-linear growth trend relative to the logarithm of output length, indicating continued benefits from test-time scaling. Additionally, we reveal that top-tier models achieve pass@32 rates exceeding 70%, indicating their potential to solve these challenging problems even if they do not consistently identify the correct reasoning path at present (see Appendix E for details). Furthermore, we conduct further analysis for reasoning efficiency, volatility, and cross-lingual robustness, providing a comprehensive view of reasoning capabilities. Notably, open-source models are narrowing the gap with proprietary commercial models over time. These analysis highlight substantial opportunities for improvement in the reasoning capabilities of future models.

The main contributions are threefold: (1) we present AMO-Bench, a novel and challenging math reasoning dataset; (2) we reveal the reasoning capability boundaries of current LLMs and highlight the potential for further improvements; and (3) beyond the accuracy score, we conduct a detailed analysis to provide deeper insights into the reasoning capabilities. We will release the data and evaluation code of AMO-Bench, hope this benchmark will facilitate further research into advancing the reasoning abilities of language models.

2 AMO-Bench

In this section, we first introduce the construction process of AMO-Bench (Section 2.1). Then, we elaborate on the grading methodology designed for AMO-Bench (Section 2.2). Finally, we perform variance estimation to characterize the volatility on AMO-Bench (Section 2.3). Figure 2 briefly illustrates the construction and grading pipeline.

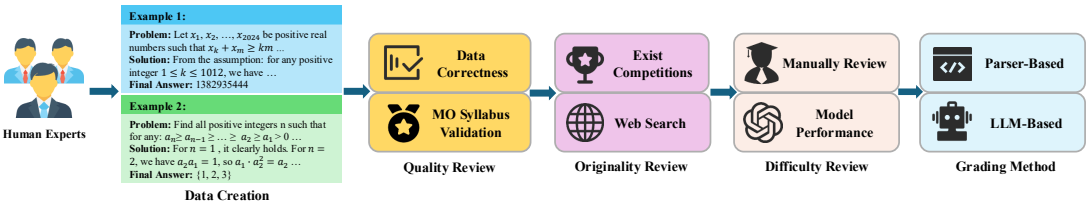


Figure 2: The construction and grading pipeline of AMO-Bench.

2.1 Construction Pipeline

To ensure the high standards of quality, originality, and difficulty level in our dataset, we have built up a comprehensive multi-stage construction pipeline that covers the entire process from question creation to final inclusion. This pipeline comprises four major stages: data creation, quality review, originality review, and difficulty review.

Data creation. All problems are independently designed by mathematics experts from top universities and educational institutions. These experts have extensive backgrounds in high school mathematics competitions, either having won MO-level mathematics competition awards or possessing experience in competition problem design. Beyond the final answer, each problem author must provide a detailed step-by-step solution. These annotated solutions will be utilized in the subsequent quality review stage and will also aid in assessing the overall difficulty of AMO-Bench (see Appendix A for dataset statistics). In addition, we provide a Chinese (CH) version of AMO-Bench to evaluate cross-lingual reasoning robustness.

Quality review. Each candidate problem undergoes blind review by at least three experts to assess its quality. This quality review stage focuses primarily on two aspects: (1) Whether the problem statement and solution are semantically unambiguous and logically correct. (2) Whether the mathematical knowledge required for the problem is within the scope typically covered in MO-level competitions such as IMO.

Originality review. The originality review stage aims to ensure that these newly created problems are not mere rewrites of publicly available materials, but demonstrate genuine originality. To this end, we assess the originality of each problem through the following methods: (1) Compare it against problems in existing datasets (e.g., AIME24/25) with 10-gram matching. (2) Conduct web searches to identify any similar online content. Additionally, during the quality review stage, experts are also required to indicate whether they have encountered highly similar questions in past competitions.

Difficulty review. To ensure that AMO-Bench presents a sufficient challenge to state-of-the-art LLMs, we implement a difficulty review stage to filter out problems lacking adequate complexity (even if they may be suitable for some MO-level competitions, e.g., the first 10 questions in AIME). Specifically, each selected problem must satisfy the following two criteria: (1) The problem must meet or exceed the IMO difficulty standards, as verified by the human expert. (2) We employed multiple advanced reasoning models (such as GPT, DeepSeek, and Gemini series models) for preliminary evaluation, requiring that at least two such models fail to correctly and consistently solve the problem[1].

2.2 Grading Method

For evaluating answers generated by LLMs, prior work has primarily utilized two approaches: parser-based grading and LLM-based grading. To fully leverage the strengths of both methods, AMO-Bench employs different grading approaches based on the specific answer type for each problem. Detailed information on the grading method and grading accuracy is provided in Appendix F.

2.3 Variance estimation

To better tackle the challenging of AMO-Bench, we run models with a relatively high temperature during inference to encourage more diverse reasoning results. However, this also increases generation stochasticity and can lead to noticeable variability in evaluation outcomes. Therefore, we conduct a variance estimation of two key metrics: (1) pairwise ranking and (2) sampling variability, to obtain more stable performance estimates (Balunović et al., 2025; Miller, 2024).

[1]For each model, our preliminary evaluation involves three samples. If all three samples are correct, the model is deemed capable of consistently solving the problem.

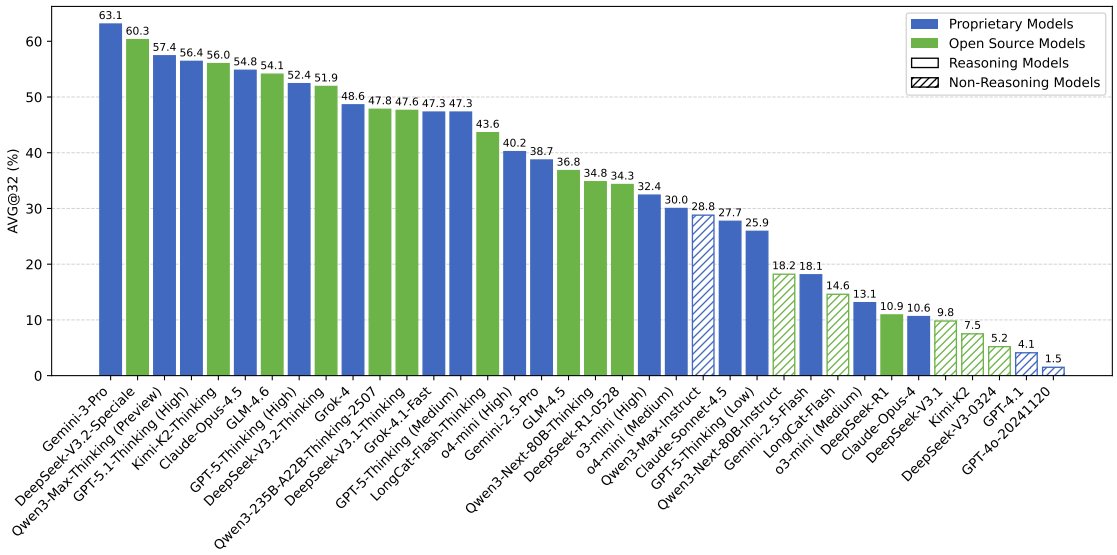


Figure 3: The AVG@32 performance of various LLMs on AMO-Bench.

For pairwise ranking. We conduct paired comparisons and compute a rank confidence interval for each model M_i to quantify ranking variance. When two models are evaluated on the same question, we treat scores as paired observations and leverage paired differences to substantially reduce the variance of the comparison, yielding more reliable significance tests and tighter confidence intervals for model ranks.

We first assess whether two models A and B differ significantly via a paired significance test. Concretely, for n independent questions, let $s_{A,i}$ and $s_{B,i}$ denote their scores on question i. We define the paired difference $s_{A-B,i} = s_{A,i} - s_{B,i}$ and the observed mean difference $\bar{s}_{A-B} = \bar{s}_A - \bar{s}_B$. Then we can estimate the paired standard error as,

$$\begin{aligned} SE_{A-B,\text{paired}} &= \sqrt{\text{Var}(s_{A-B})/n} \\ &= \sqrt{\frac{1}{n(n-1)} \sum_i \left(s_{A-B,i} - \bar{s}_{A-B}\right)^2}, \end{aligned} \tag{1}$$

where $\text{Var}(\cdot)$ represents the variance function. Under a normal approximation, we construct a 95% confidence interval $[\bar{d} \pm 1.96 \cdot SE_{A-B,\text{paired}}]$ and use it for significance testing: if the interval excludes 0, we conclude that models A and B differ significantly; otherwise the difference is not statistically significant.

Subsequently, we perform pairwise comparisons to every other model for each model M_i. Let $N_{\text{better}}(i)$ denote the number of models that are significantly better, and $N_{\text{worse}}(i)$ denote the number of models that are significantly worse. This yields the lower and upper bounds of the rank confidence interval,

$$\begin{aligned} r_i^{\text{lower}} &= 1 + N_{\text{better}}(i), \\ r_i^{\text{upper}} &= N - N_{\text{worse}}(i). \end{aligned} \tag{2}$$

Given a total of N models, the rank confidence interval for model M_i is $[r_i^{\text{lower}}, r_i^{\text{upper}}]$.

For sampling variability. To quantify sampling variability under AVG@k (e.g., AVG@32), we estimate a 95% confidence interval via Monte Carlo simulation.

Specifically, for a model evaluated on n questions with per-question success probabilities $\{p_i\}_{i=1}^n$, we first conduct the following sampling process,

$$C_i \sim \text{Binomial}(k, p_i), \qquad i = 1, \ldots, n, \tag{3}$$

and then compute a simulated AVG@k score,

$$S = \frac{1}{n} \sum_{i=1}^{n} \frac{C_i}{k}. \tag{4}$$

We repeat this sampling process for T times and yield $\{S^{(t)}\}_{t=1}^T$. Therefore, the 95% Confidence Interval is constructed from the empirical quantiles of the simulated score distribution,

$$\begin{aligned} &\text{CI}_{0.95}(S) \\ &= \left[Q_{0.025}\left(\{S^{(t)}\}\right),\ Q_{0.975}\left(\{S^{(t)}\}\right)\right], \end{aligned} \tag{5}$$

where $Q_q(\cdot)$ denotes the empirical q-th quantile.

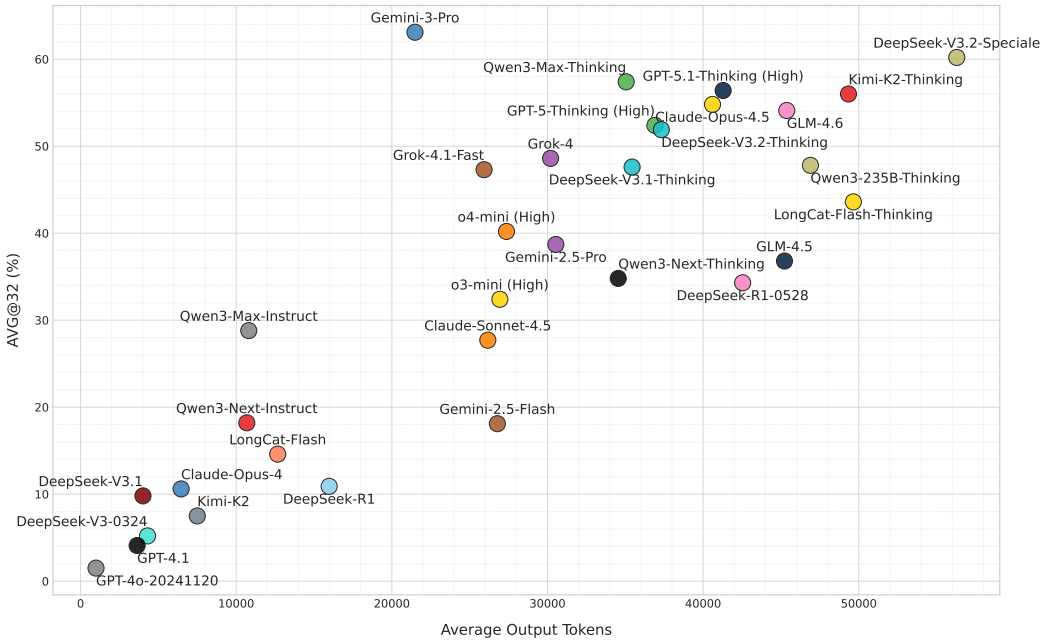


Figure 4: The AVG@32 performance of LLMs vs. the average model output length.

Finally, the sampling variability interval is characterized by the mean score $\mathbb{E}[S]$ and half-width error ϵ, (i.e., $[\mathbb{E}[S] - \epsilon,\ \mathbb{E}[S] + \epsilon]$).

$$\mathbb{E}[S] = \frac{1}{n}\sum_{i=1}^{n} p_i, \qquad \epsilon = \big(Q_{0.975} - Q_{0.025}\big)/2. \tag{6}$$

3 Experiments

In this section, we present the experimental results on AMO-Bench[2]. We first describe the experimental setup (Section 3.1), followed by a discussion of the main results and analysis (Section 3.2).

3.1 Experimental Setup

Models. To conduct a comprehensive and representative evaluation on AMO-Bench, we select a diverse set of leading LLMs, encompassing both open-source models and proprietary models. Specifically, the evaluation includes top-tier models provided by OpenAI (OpenAI, 2025), Gemini (Gemini Team, 2025), Anthropic (Anthropic, 2025), DeepSeek (Guo et al., 2025), Qwen (Yang et al., 2025), GLM (GLM-4.5 Team, 2025), Moonshot (Kimi Team, 2025), and LongCat (Meituan LongCat Team, 2025b). In addition to evaluating reasoning models that have been specifically enhanced for long-term thinking tasks, we also incorporated several powerful non-reasoning models to demonstrate their potential in tackling complex reasoning challenges.

Sampling settings. We set the `temperature` of sampling to 1.0 for reasoning models and 0.7 for non-reasoning models. For all evaluated models, we use `top-k`=50 and `top-p`=0.95 during sampling. We configure the maximum context/output length to the highest allowable limit for each model during inference. This avoids underestimating the reasoning capabilities of the model due to restrictions on the token budget. To ensure the stability of the final evaluation results, we sampled the results from each model 32 times and reported the average performance of these 32 results as the final metric (denoted as AVG@32). Appendix D illustrates the fluctuation of the average result across different sampling times.

3.2 Results and Analysis

AMO-Bench presents a significant challenge. Figure 3 presents the AVG@32 performance of various leading LLMs, categorized by proprietary/open-source status and reasoning/non-reasoning properties. Overall, all these models still struggle with the significant challenges presented by AMO-Bench. Even the highest

[2]To facilitate easier reproduction and utilization of AMO-Bench, you can take a fast try on the AMO-Bench-P subset, which includes only the 39 parser-based grading problems from AMO-Bench. Appendix G presents the AVG@32 performance of LLMs on AMO-Bench-P.

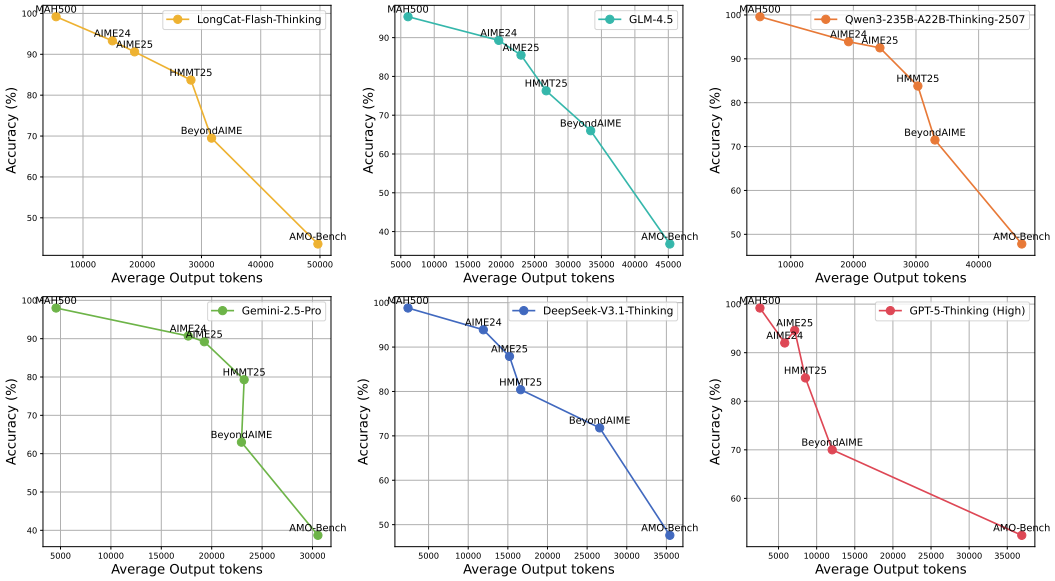


Figure 5: The relationship between accuracy and average output length on different math benchmarks.

performing model Gemini-3-Pro reaches just 63.1%, while most others score below 50%. This indicates substantial room for improvement in complex reasoning abilities across all current language models. Moreover, both proprietary and open-source reasoning models occupy top ranks in the leaderboard, indicating that recent open-source advancements are closing the gap with leading commercial models. The best-performing open-source model is only about 3% lower than the top proprietary result. Besides reasoning models, some non-reasoning models demonstrate a performance exceeding expectations, such as Qwen3-Max-Instruct and LongCat-Flash. These non-reasoning models even outperforms several reasoning models such as o3-mini (Medium), indicating their significant potential in tackling complex reasoning tasks. Beyond the main results outlined above, we also provide further analysis and insights based on the AMO-Bench experimental findings.

Comparison of reasoning efficiency. Figure 4 shows the average output length and the AVG@32 performance of each model. Overall, it demonstrates a clear trend that higher-performing models tend to require more output tokens. Most of models that reach higher than 50% AVG@32 scores utilize more than 35K completion tokens. Even among non-reasoning models, those with superior performance are distinguished by their ability to process more tokens, sometimes reaching levels comparable to reasoning models. Additionally, when examining models within the same series, there are notable improvements in reasoning efficiency over time. For example, o4-mini (High) outperforms o3-mini (High) at similar or slightly increased token counts. Likewise, Gemini-3-Pro shows significant gains compared to Gemini-2.5-Pro with even significantly less output tokens.

The model output length could indicate the reasoning challenge of the benchmark. Appendix A provides a pre-analysis of benchmark difficulty based on annotated solution lengths. Here, we offer a post-hoc analysis of benchmark difficulty based on the relationship between model performance and model output length. Figure 5 clearly demonstrates that the average output length of each model increases as the reasoning benchmark becomes more challenging. Specifically, across six models, benchmarks with higher accuracy scores (such as MAH500 and AIME24) correspond to shorter average outputs, while those with lower scores (like AMO-Bench) require significantly longer responses. This suggests that harder benchmarks demand more elaborate reasoning steps or explanations from the models, resulting in increased token usage. These results demonstrate that the model output length could be an indicator of reasoning challenge in the benchmark.

Performance on AMO-Bench still benefits from test-time scaling. The reasoning efficiency results discussed above indicate a correlation between model performance and output length. Here,

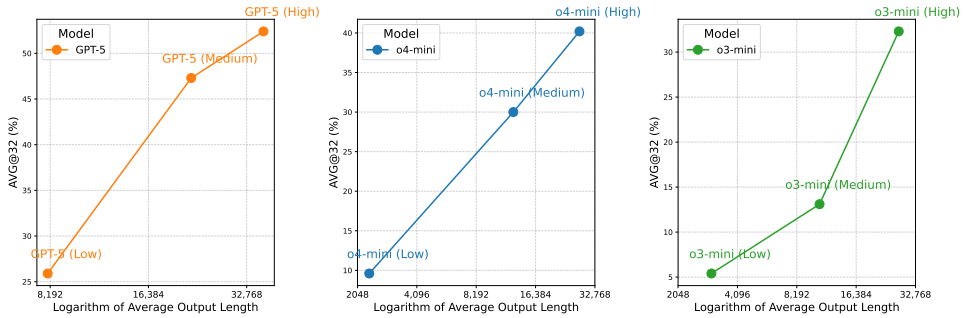


Figure 6: The model performance and output length under different reasoning effort settings.

we conduct a more rigorous analysis by directly controlling the reasoning effort for the same model. As shown in the Figure 6, all three models (GPT-5 o4-mini and o3-mini) exhibit a near-linear growth trend in AVG@32 as the logarithm of average output length increases. Such a trend is highly aligned with earlier experimental observations from existing benchmarks such as MATH500 and AIME24 (Muennighoff et al., 2025). This indicates that further increasing the inference budget will further drive improvements on AMO-Bench.

Volatility Analysis on AMO-Bench. To robustly assess performance differences on AMO-Bench, we follow Section 2.3 and quantify volatility from two perspectives: (1) pairwise ranking and (2) sampling variability of AVG@32. Table 1 shows the pairwise ranking intervals and score, while Table 2 summarizes the sampling variability for AVG@32. Overall, the intervals are sufficiently tight to separate a large portion of models, indicating AMO-Bench provides a reliable basis for comparative evaluation despite its limited number of questions. Notably, Gemini-3-Pro, DeepSeek-V3.2-Speciale and Qwen3-Max-Thinking form a clear top tier, while lower-performing models remain well below 20% AVG@32.

Cross-lingual robustness remains challenging. Cross-lingual performance provides important insight into whether a model's reasoning ability generalizes beyond a single language. As shown in Figure 9, we conduct a contrast analysis of model performance on AMO-Bench in both English (EN) and Chinese (CH). Although top-tier reasoning models exhibit strong cross-lingual consistency, several models still suffer substantial performance drops on the CH benchmark. This underscores that cross-lingual transfer of reasoning remains challenging for current models, indicating potential room for further improvement.

Table 1: Pairwise ranking volatility on AMO-Bench.

Model	Rank	Acc (%)
Gemini-3-Pro	1–4	63.1
DeepSeek-V3.2-Speciale	1–7	60.3
Qwen3-Max-Thinking (Preview)	1–9	57.4
GPT-5.1-Thinking (High)	1–13	56.4
Kimi-K2-Thinking	2–12	56.0
Claude-Opus-4.5	2–12	54.8
GLM-4.6	2–13	54.1
GPT-5-Thinking (High)	3–13	52.4
DeepSeek-V3.2-Thinking	3–13	51.9
Grok-4	4–15	48.6
Qwen3-235B-A22B-Thinking-2507	4–16	47.8
DeepSeek-V3.1-Thinking	6–16	47.6
Grok-4.1-Fast	4–16	47.3
LongCat-Flash-Thinking	10–16	43.6
o4-mini (High)	11–19	40.2
Gemini-2.5-Pro	10–21	38.7
GLM-4.5	15–20	36.8
Qwen3-Next-80B-Thinking	15–22	34.8
DeepSeek-R1-0528	15–21	34.3
o3-mini (High)	16–22	32.4
Qwen3-Max-Instruct	17–22	28.8
Claude-Sonnet-4.5	19–22	27.7
Qwen3-Next-80B-Instruct	23–24	18.2
Gemini-2.5-Flash	23–25	18.1
LongCat-Flash	24–27	14.6
DeepSeek-R1	25–29	10.9
Claude-Opus-4	25–30	10.6
DeepSeek-V3.1	26–29	9.8
Kimi-K2	26–30	7.5
DeepSeek-V3-0324	28–32	5.2
GPT-4.1	30–32	4.1
GPT-4o-20241120	30–32	1.5

The gap between open-source and proprietary models on AMO-Bench is narrowing. Figure 7 illustrates model performances relative to release dates, highlighting the SOTA open-source and proprietary models across different periods. It reveals a clear upward trend in top-tier performance; however, AMO-Bench remains far from saturated. Furthermore, recent advancements in open-source models are rapidly narrowing the gap with proprietary counterparts, with the leading models now performing within a 5% accuracy gap.

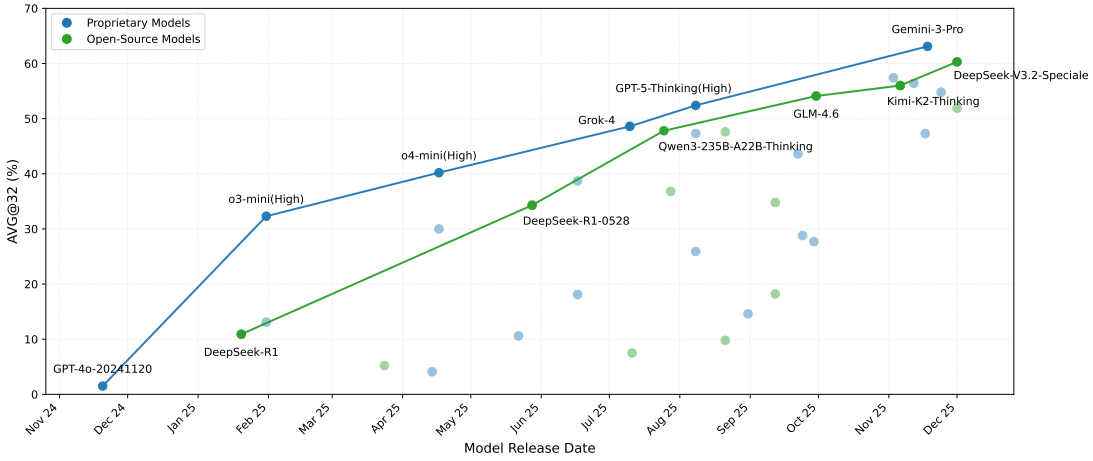


Figure 7: Comparison of proprietary and open-source models' performance on AMO-Bench over time.

4 Related Work

Evaluating LLMs on mathematical problems has been a critical aspect of assessing reasoning capabilities. Early datasets such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) provided initial explorations to evaluate these abilities. However, model performance on these benchmarks has quickly reached saturation. To further advance the study of mathematical proficiency in LLMs, recent work has shifted toward more challenging benchmarks.

In terms of increasing difficulty, two primary lines of work have emerged. One line focuses on Mathematical Olympiad (MO)-level problems, which rely on a specific range of math knowledge and require complex and intuitive reasoning skills. For instance, Omni-MATH (Gao et al., 2024) introduces a multi-subject evaluation suite designed to rigorously test mathematical reasoning and generalization; OlymMATH (Sun et al., 2025) collects MO-level problems from printed publications and evaluates mathematical reasoning by offering problems of two difficulty levels; MathOdyssey (Fang et al., 2025) broadens the scope to include more complex tasks, with a particular focus on long-range and compositional reasoning; BeyondAIME (ByteDance-Seed, 2025) collects problems similar in style to AIME with increased difficulty and expanded data scale; MathArena (Balunović et al., 2025) rapidly tracks model performance in newly held MO-level competitions and explores evaluation paradigms for proof-based competitions such as the IMO and USAMO. Our proposed AMO-Bench also falls within this category and it stands as one of the most challenging benchmarks at the time of writing.

The other line of work focuses on problems derived from graduate-level examinations or advanced mathematical research. For instance, RealMath (Zhang et al., 2025) provides a comprehensive evaluation of LLMs in real-world mathematical tasks, assessing their reasoning capabilities across a diverse range of research-level content; FrontierMath (Glazer et al., 2024) covers computationally intensive problems and abstract questions across most branches of mathematics, highlighting the significant gap between LLMs and the prowess of the mathematical community; HARDMath2 (Roggeveen et al., 2025) focuses on approximation-based mathematical problems, particularly those commonly encountered in applied sciences and engineering; HLE (Phan et al., 2025) constructs a final closed-ended academic benchmark spanning multiple subjects, evaluating reasoning capabilities on human frontier knowledge. Beside requiring the reasoning abilities, these datasets also challenge models by demanding extensive and deep mathematical knowledge.

5 Conclusion

We introduce AMO-Bench, an advanced mathematical reasoning benchmark featuring problems at the IMO-level difficulty or higher, consisting of 50 human-crafted questions. Experimental results across various LLMs demonstrate that contemporary LLMs still struggle with the significant challenges presented by AMO-Bench. Despite these limited performances, our further analysis underscore substantial opportunities for advancing mathematical reasoning capabilities in current LLMs.

Limitations

This work introduced AMO-Bench to assess mathematical reasoning capabilities in current LLMs. The main limitation of this work stems from the inherent variability within the evaluation process. (1) Due to the limited data size of the AMO-Bench and the requirement for high sampling temperatures in the inference process of the reasoning models, the final evaluation results are inherently influenced by randomness and sampling noise. To alleviate the issue, we conduct variance analyses to explicitly quantify volatility and reduce over-interpretation of small differences. (2) While we adopt a hybrid grading method, using parser-based grading for structured answers and LLM-based grading for descriptive responses, each approach entails inherent trade-offs. As a result, there remains room to further improve both grading accuracy and computational efficiency. Overall, we hope the future work will expand the benchmark scale and strengthen automated verification to further improve the stability, fidelity, and scalability of evaluations on AMO-Bench.

Ethics Statement

AMO-Bench evaluates mathematical reasoning with competition-style problems and final-answer grading, and it is not intended to elicit sensitive or harmful content. Accordingly, we do not anticipate substantial direct ethical risks from the benchmark itself. Nevertheless, many reasoning models produce lengthy intermediate traces during inference. Although such traces are not required for evaluation, they may inadvertently contain inappropriate assumptions or other content unrelated to the task, introducing implicit policy and safety concerns. To promote more reliable and transparent model behavior, AMO-Bench provides human-annotated reasoning paths for each problem as a reference signal. These expert-written solutions facilitate prompt development and systematic error analysis, and can also be used to monitor unreasonable phenomena during the model thinking process. We also call upon the research community to undertake further studies into the monitoring of the model reasoning processes.

References

Anthropic. 2025. System card: Claude opus 4 and claude sonnet 4.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. Matharena: Evaluating llms on uncontaminated math competitions. arXiv preprint arXiv:2505.23281.

ByteDance-Seed. 2025. Beyondaime: Advancing math reasoning evaluation beyond high school olympiads.

ByteDance Seed. 2025. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. Preprint, arXiv:2504.13914.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

DeepSeek-AI. 2025. Deepseek-v3 technical report. Preprint, arXiv:2412.19437.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2025. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. Scientific Data, 12(1):1392.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, and 1 others. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. In The Thirteenth International Conference on Learning Representations.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, arXiv:2507.06261.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. arXiv preprint arXiv:2411.04872.

GLM-4.5 Team. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. Preprint, arXiv:2508.06471.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature, 645(8081):633–638.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 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.

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 Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Kimi Team. 2025. Kimi k2: Open agentic intelligence. Preprint, arXiv:2507.20534.

Meituan LongCat Team. 2025a. Longcat-flash technical report. Preprint, arXiv:2509.01322.

Meituan LongCat Team. 2025b. Longcat-flash-thinking technical report. Preprint, arXiv:2509.18883.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations. Preprint, arXiv:2411.00640.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In Workshop on Reasoning and Planning for Large Language Models.

OpenAI. 2024. Openai o1 system card. Preprint, arXiv:2412.16720.

OpenAI. 2025. Gpt-5 system card.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. 2025. Proof or bluff? evaluating llms on 2025 usa math olympiad. Preprint, arXiv:2503.21934.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. arXiv preprint arXiv:2501.14249.

James V Roggeveen, Erik Y Wang, Will Flintoft, Peter Donets, Lucy S Nathwani, Nickholas Gutierrez, David Ettel, Anton Marius Graf, Siddharth Dandavate, Arjun Nageswaran, and 1 others. 2025. Hardmath2: A benchmark for applied mathematics built by students as part of a graduate class. arXiv preprint arXiv:2505.11774.

Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models. arXiv preprint arXiv:2503.21380.

Tencent Hunyuan Team. 2025. Hunyuan-turbos: Advancing large language models through mamba-transformer synergy and adaptive chain-of-thought. Preprint, arXiv:2505.15431.

xAI. 2025. Grok 4 model card.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. 2025. Realmath: A continuous benchmark for evaluating language models on research-level mathematics. arXiv preprint arXiv:2505.12575.

A Dataset Statistics

Problem categories. Referring several official competition syllabus, we categorize the 50 problems of AMO-Bench into the following five primary categories: Algebraic Equations & Inequalities (11/50), Functions & Sequences (13/50), Geometry (5/50), Number Theory (9/50), and Combinatorics (12/50). Figure 8a show the overall distribution of problem categories in AMO-Bench.

Length distribution of human-annotated solutions. Since the problems in our AMO-Bench are equipped with manually annotated solutions, we can preliminarily analyze the reasoning complexity of these problems from the view of solution length. We measure solution length in terms of token count[3]. Additionally, we compare the distribution of solution lengths with those from AIME24[4] and MATH500[5]. Figure 8b illustrates the solution length distributions across these benchmarks. It reveals that solutions in AMO-Bench exhibit significantly higher lengths, indicating that problems in this benchmark are inherently more challenging and require more complex reasoning to arrive at the final answer. We conduct a further analysis of the model solution lengths in Section 3.2.

B Prompt Templates

Query prompt template. In order to guide LLMs in generating answers in a parser-readable format, we use the following prompt template guide the model generation(e.g., Example 1). There are mainly three requirements in the instruction: the answer prefix (i.e., `### The final answer is:`), the LaTeX box environment (i.e., `\boxed{}`), and the precision requirement (at least four decimal places).

Grading prompt template. We employ the LLM-based grading using o4-mini (Low) as the grading model, and use the following grading prompt to verify the equivalence between the LLM output and the reference answer(e.g., Example 2). The template frames the task as a strict equivalence check and requires a binary decision (i.e., `Correct` / `Incorrect`) in a fixed output format.

[3]We use the tokenizer of DeepSeek-V3.1 model to count tokens in solutions.

[4]https://huggingface.co/datasets/HuggingFaceH4/aime_2024.

[5]https://huggingface.co/datasets/HuggingFaceH4/MATH-500.

C More Details About Dataset Creation Process

In Section 2.1, we have introduced the dataset creation pipeline and the major principles for data collection and filtering. Here, we provide further details in the dataset creation process. We gather data from collaborating math competition experts through surveys and email contributions. Experts interested in contributing are briefed on specifications such as problem type, originality, and difficulty. We pay contributors and referrers a fee of \$150–\$200 for every problem that passes our final check. Typically, math competition problems involve no ethical or moral concerns. To further guarantee this, we submitted the final dataset and the grading algorithm code to our internal data risk oversight team, and the submission successfully passed the review.

D Analysis of AVG@k

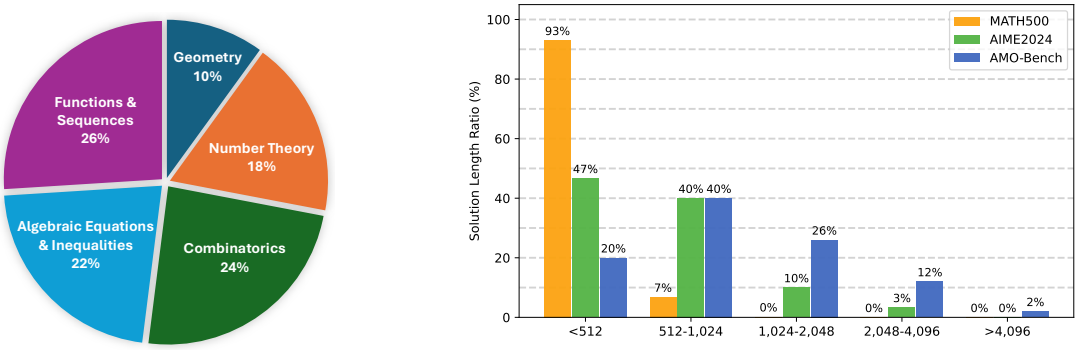
Increasing the sample size k yields more stable performance estimates on AMO-Bench. Figure 10 illustrates the fluctuation of the average performance across different sampling times. It shows that as the sampling time grows, the models' performance become more stable. When sampling 32 times, the average model performance exhibits a relatively small fluctuation and rarely appears to reverse the model ranking order (the reverse-order phenomenon).

E Analysis of Pass@k

Top-tier models demonstrate promising potential for improvement on AMO-Bench. Existing work reveals that the pass@k performance of the model can reflect its inherent potential to achieve further improvement through reinforcement learning. Inspired by this, we illustrate the pass@k of evaluated models to indicate their inner potential. As shown in Figure 11, the pass@k metric exhibits rapid growth as k increases from 1 to 8, followed by a sustained but gradual improvement as k continues to rise. Notably, the top-tier reasoning models achieve over 70% performance on the pass@32 metric. These results highlight the significant room for improvement in the reasoning capabilities of LLMs.

F Grading details

Grading method. For evaluating answers generated by LLMs, prior work has primarily utilized



(a) Distribution of problem categories.

(b) Comparison of solution lengths.

Figure 8: Basic statistics of AMO-Bench. (a) The distribution of problem categories in AMO-Bench. (b) The distribution of human-annotated solutions in AMO-Bench as well as the comparison with MATH500 and AIME24.

two approaches: parser-based grading and LLM-based grading. Parser-based grading offers high efficiency and accuracy when the model's response can be successfully parsed; however, its applicability is limited to simple answer formats such as numerical values or sets, making it challenging to assess more complex answers. In contrast, LLM-based grading provides greater flexibility across diverse answer types but may be less efficient and does not consistently guarantee accuracy.

To fully leverage the strengths of both grading methods, AMO-Bench employs different grading approaches based on the specific answer type for each problem. Specifically, problems in AMO-Bench are divided into four main answer types: numerical answers (e.g., Example 3), set answers (e.g., Example 4), variable-expression answers (e.g., Example 5 which requires providing the general formula for an arithmetic sequence), and descriptive answers (e.g., Example 6 which involves comprehensively considering multiple scenarios). The prompt templates for used for grading are contained in Appendix B.

For problems requiring numerical, set, or variable-expression answers (39 out of 50), we employ the parser-based grading. The evaluated LLMs are instructed to format their final responses as `\boxed{<answer>}`. We then utilize the tools provided by `math-verify`[6] to parse these answers and verify the equivalence with the ground truth. Moreover, if the model answer containing decimal values, we require an accuracy of at least four decimal places. For variable-expression answers, we assign multiple sets of values to the variables in the expression, then verify whether the values of the generated expression match that of the ground-truth expression. We also manually review the parsing results during the preliminary evaluation and adjust the post-processing algorithms.

For problems requiring descriptive answers (11 out of 50), we use LLM-based grading with o4-mini (Low) serving as the grading model. To ensure robust assessment, majority voting is performed across five independent grading samples for each response. Additionally, during preliminary evaluation, we manually verify the correctness of LLM-based grades for all descriptive answers and revise answer descriptions where needed to enhance grading accuracy.

Grading accuracy. Prior to conducting the large-scale evaluation, we performed a manual quality check to ensure the reliability of the designed grading method. This assessment included 1,000 responses in total generated by 10 different LLMs (across English and Chinese). The results indicate that the grading accuracy reached 99.2%, providing strong validation for the effectiveness of the grading method on AMO-Bench.

G Performance on AMO-Bench-P Subset

To facilitate easier reproduction and use of AMO-Bench, you can utilize the AMO-Bench-P subset, which includes only the 39 parser-based grading problems from AMO-Bench. Table 2 also presents the AVG@32 performance of LLMs on AMO-Bench-P. In general, performance on AMO-Bench-P tends to be slightly higher than on the full AMO-Bench, as problems requiring complex descriptive answers are inherently more challenging than those with simple-format answers.

[6] https://github.com/huggingface/Math-Verify.

Example 1: Query Prompt Template

...
After solving the above problem, please output your final answer in the following format:
The final answer is: $\boxed{<your answer>}$
Example:
The final answer is: $\boxed{123}$
The final answer should be given as precisely as possible (using LaTeX symbols such as \sqrt, \frac, \pi, etc.). If the final answer involves a decimal approximation, it must be accurate to at least four decimal places.

Example 2: Grading Prompt Template

For the following math problem, we have the reference answer and the student's answer.
Determine whether the student's answer is equivalent to the reference answer.
If equivalent, output "Correct".
If not equivalent, output "Incorrect".

Problem
...

Reference Answer
...

Student Answer
...

Now, please provide your judgment.
Please strictly follow the format below to summarize your conclusion at the end of your judgment:
Conclusion: Correct/Incorrect
If the answer involves a decimal approximation, it must be accurate to at least four decimal places.

Example 3: Problem with Numerical Answer

Question: Let $x_1, x_2, \cdots, x_{2024}$ be positive real numbers such that $x_k + x_m \geq km$ for any $1 \leq k < m \leq 2024$. Find the minimum value of $x_1 + x_2 + \cdots + x_{2024}$.
Answer: $\boxed{1382935444}$

Example 4: Problem with Set Answer

Question: Find all positive integers **n** such that for any: $a_n \geq a_{n-1} \geq a_{n-2} \geq \cdots\cdots a_2 \geq a_1 > 0$, satisfying $\sum_{k=1}^{n} a_k = \sum_{k=1}^{n} \frac{1}{a_k}$, the inequality $\prod_{k=1}^{n} a_k^k \geq 1$ holds.
Answer: $\boxed{\{1, 2, 3\}}$

Example 5: Problem with Variable-Expression Answer

Question: The sequence $\{a_n\}_{n=1}^{\infty}$ consists of positive terms, with $a_1 = 7$, $a_2 = 2$, and satisfies the recurrence relation

$$8a_{n+2}^4 = 3 + 4a_{n+1} + a_n \quad (n \in \mathbb{N}^*).$$

Find the general term formula for this sequence.
Answer:

$$\boxed{\frac{(2+\sqrt{3})^{2^{2-n}} + (2-\sqrt{3})^{2^{2-n}}}{2}}$$

Example 6: Problem with Descriptive Answer

Question: Let n be an integer with $n > 2$. Real numbers $a_1, a_2, \ldots, a_n$ satisfy

$$\sum_{k=1}^{n} a_k = 2n, \qquad \sum_{k=1}^{n} k\,|a_k| = 4n.$$

Find the minimum value of $a_1^2 + a_2^2 + \cdots + a_n^2$.
Answer: For $n = 3$, the minimum of $a_1^2 + a_2^2 + a_3^2$ is 12.

For $n \geq 4$, the minimum of $a_1^2 + a_2^2 + \cdots + a_n^2$ is $\frac{6n^2}{5}$.

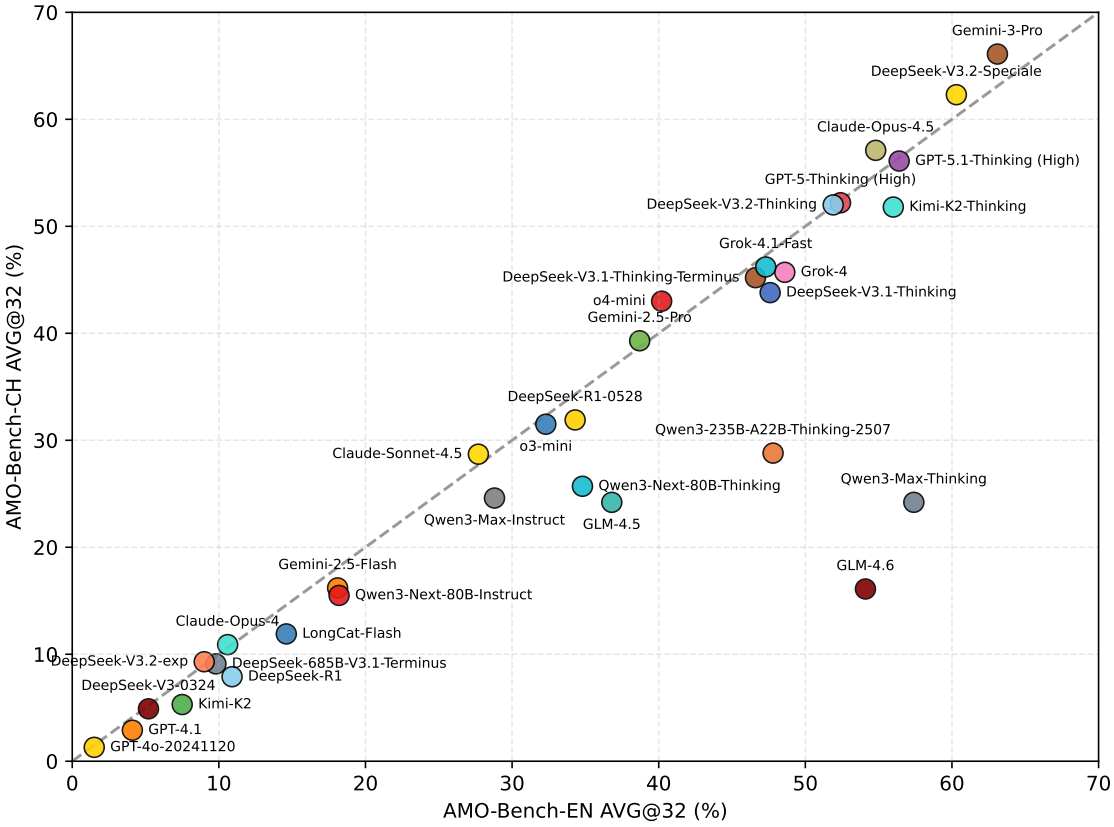


Figure 9: The AVG@32 performance of LLMs on AMO-Bench in English (EN) and Chinese (CH).

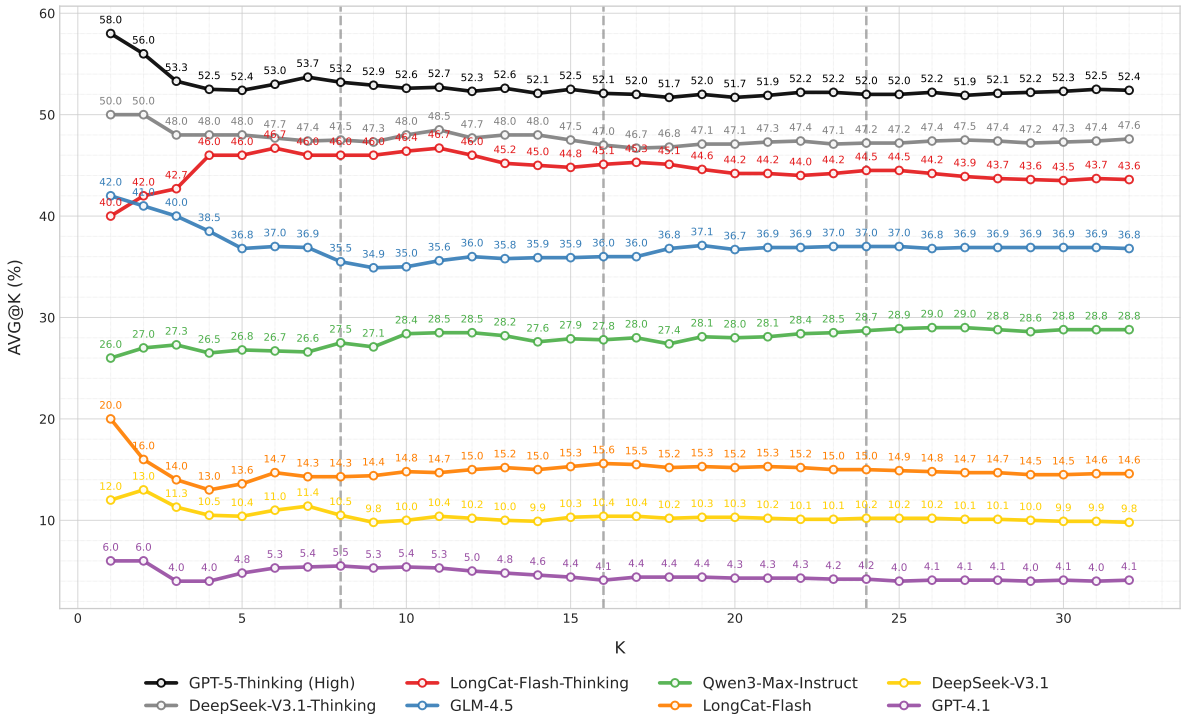


Figure 10: The AVG@k trend of various LLMs with increasing k.

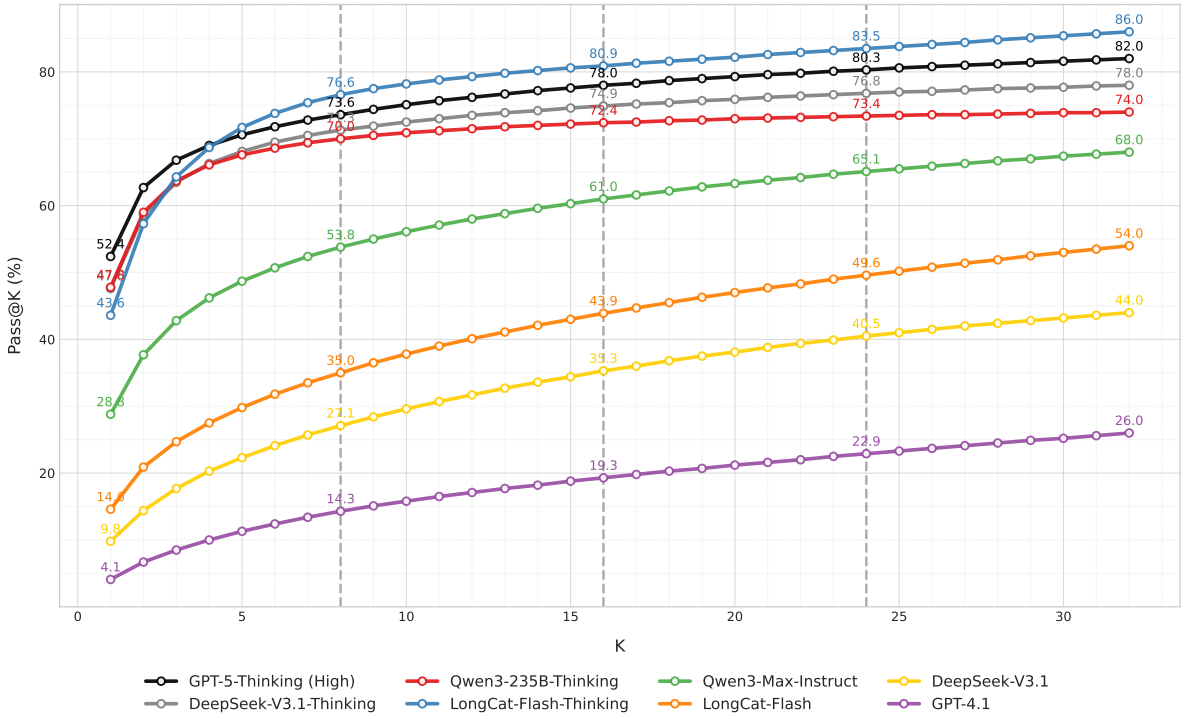


Figure 11: The the pass@k trend of various LLMs with increasing k.

Table 2: The AVG@32 performance of LLMs on AMO-Bench and its subset AMO-Bench-P, which includes only the 39 parser-gradable problems. We show the AVG@32 with the 95% confidence intervals.

Model	AMO-Bench (%)	AMO-Bench-P (%)
Gemini-3-Pro	63.1 ± 1.4	67.4 ± 1.7
DeepSeek-V3.2-Speciale	60.3 ± 1.2	62.3 ± 1.4
Qwen3-Max-Thinking (Preview)	57.4 ± 1.5	60.0 ± 1.7
GPT-5.1-Thinking (High)	56.4 ± 1.3	58.9 ± 1.6
Kimi-K2-Thinking	56.0 ± 1.6	55.7 ± 1.8
Claude-Opus-4.5	54.8 ± 1.6	59.8 ± 1.8
GLM-4.6	54.1 ± 1.6	55.4 ± 1.8
GPT-5-Thinking (High)	52.4 ± 1.6	54.8 ± 1.8
DeepSeek-V3.2-Thinking	51.9 ± 1.3	53.2 ± 1.5
Grok-4	48.6 ± 1.5	55.1 ± 1.8
Qwen3-235B-A22B-Thinking-2507	47.8 ± 1.6	56.2 ± 2.0
DeepSeek-V3.1-Thinking	47.6 ± 1.6	53.0 ± 1.9
Grok-4.1-Fast	47.3 ± 1.3	55.1 ± 1.6
LongCat-Flash-Thinking	43.6 ± 1.8	45.3 ± 2.0
o4-mini (High)	40.2 ± 1.7	43.8 ± 2.0
Gemini-2.5-Pro	38.7 ± 1.6	41.7 ± 1.9
GLM-4.5	36.8 ± 1.6	41.0 ± 1.9
Qwen3-Next-80B-Thinking	34.8 ± 1.5	37.4 ± 1.9
DeepSeek-R1-0528	34.3 ± 1.7	37.1 ± 2.0
o3-mini (High)	32.4 ± 1.5	34.0 ± 1.7
Qwen3-Max-Instruct	28.8 ± 1.4	30.9 ± 1.7
Claude-Sonnet-4.5	27.7 ± 1.6	29.5 ± 1.8
Qwen3-Next-80B-Instruct	18.2 ± 1.3	17.8 ± 1.5
Gemini-2.5-Flash	18.1 ± 1.5	18.0 ± 1.7
LongCat-Flash	14.6 ± 1.2	14.9 ± 1.4
DeepSeek-R1	10.9 ± 1.1	11.7 ± 1.4
Claude-Opus-4	10.6 ± 1.1	11.4 ± 1.3
DeepSeek-V3.1	9.8 ± 1.1	9.6 ± 1.3
Kimi-K2	7.5 ± 1.0	8.4 ± 1.2
DeepSeek-V3-0324	5.2 ± 0.9	5.4 ± 1.0
GPT-4.1	4.1 ± 0.8	4.8 ± 0.9
GPT-4o-20241120	1.5 ± 0.5	1.9 ± 0.6