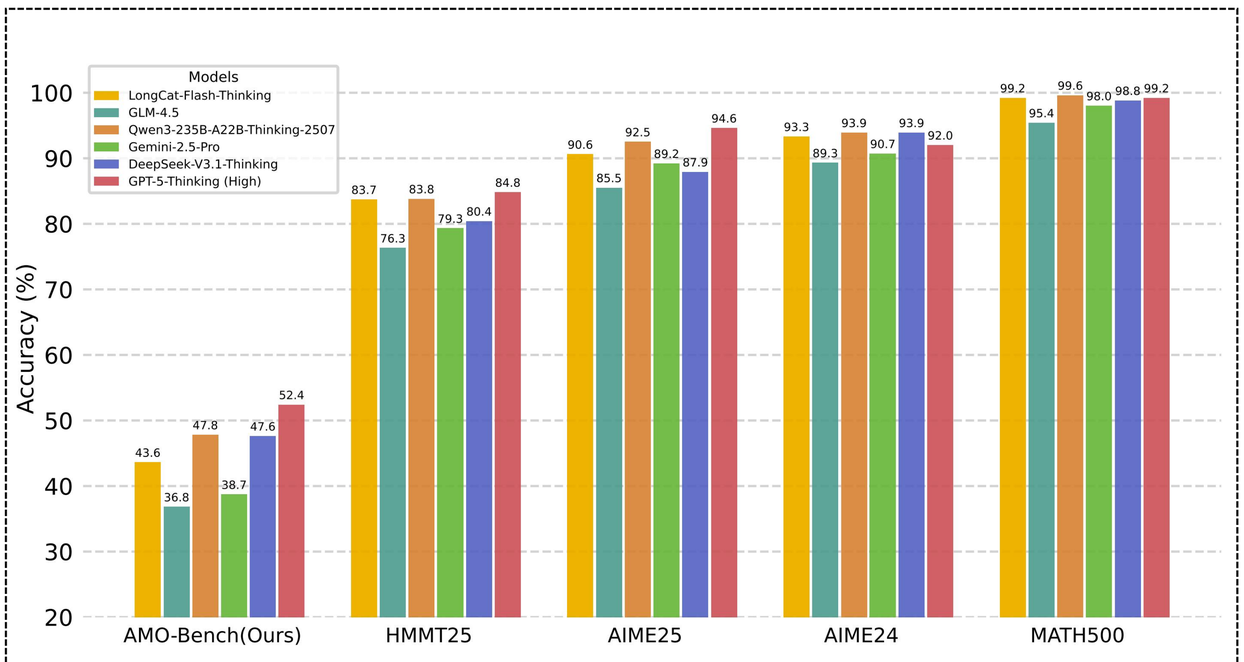
AMO-Bench: Large Language Models Still Struggle in High School Math Competitions

ACL 2026 SAN DIEGO

Junlin Liu[1], Shengnan An[2], Shuang Zhou[2], Dan Ma[2], Yehao Lin[2], Xinxuan Lv[2], Xuanlin Wang[3], Xiaoyu Li[2], Ziwen Wang[2], Xuezhi Cao[2], Xunliang Cai[2]

[1] University of Chinese Academy of Sciences
[2] Meituan
[3] Harbin Institute of Technology

1.Motivations and Features



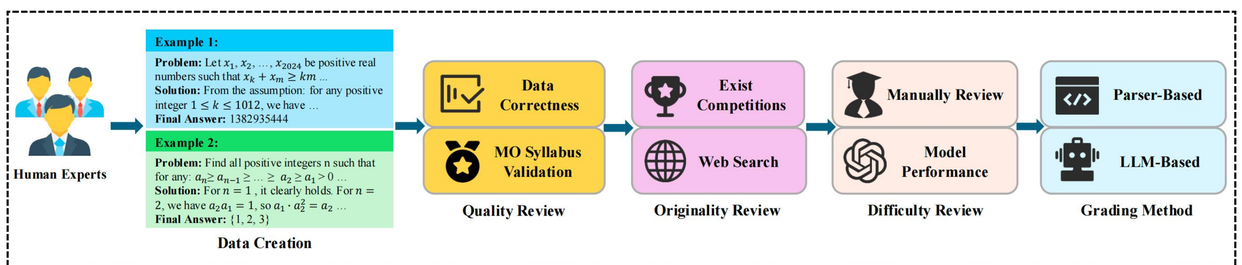
Challenges

(1) Performance saturation.
(2) Data memorization from competitions.
(3) Proof–based problems require expert verification.

Key Features

(1) Original problems.
(2) Guaranteed difficulty.
(3) Finalanswer based grading.
(4) Humanannotated reasoning paths.

2.Construction Pipeline



(1) Data creation: All 50 problems are newly crafted by human experts to prevent data leakage from existing resources.
(2) Quality review: Cross–validated by experts to meet IMO standards. LLM–based filtering ensures problems challenge SOTA models.
(3) Originality review: Enables efficient automated grading via parser–based or LLM–based methods, balancing cost and generalizability.
(4) Difficulty review.: Expert–written step–by–step solutions for each problem, supporting error analysis and prompt engineering research.

3.Variance Estimation

(1) For pairwise ranking: We conduct paired comparisons and compute a rank confidence interval for each model to quantify ranking variance. Given a total of N models, the rank confidence interval for model is:

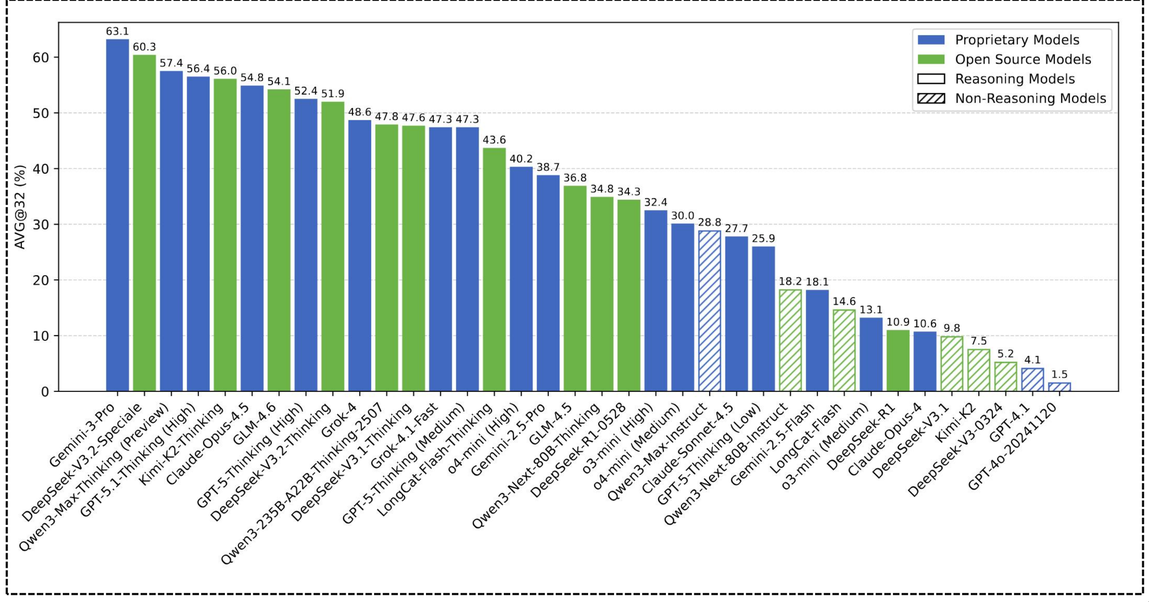
$$r_i^{\text{lower}} = 1 + N_{\text{better}}(i), r_i^{\text{upper}} = N - N_{\text{worse}}(i).$$

(2) For sampling variability: we estimate a 95% confidence interval via Monte Carlo simulation.

$$\text{CI}_{0.95}(S) = \left[Q_{0.025}\left(\{S^{(t)}\}\right),\ Q_{0.975}\left(\{S^{(t)}\}\right) \right],$$

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

4.Leaderboard



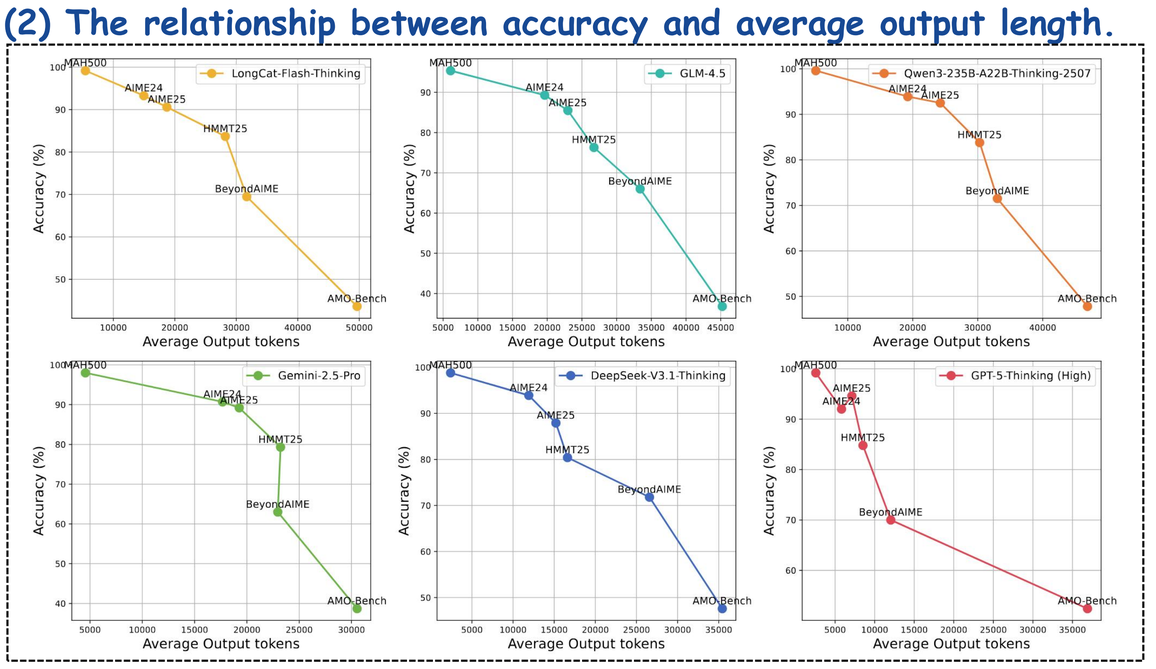
5.Results and Analysis

(1) The performance and variance of LLMs 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

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

(2) The relationship between accuracy and average output length.



(3) Comparison of reasoning efficiency.