Reasoning 2

Haw-Shiuan Chang

Deadlines

- **https://people.cs.umass.edu/~hschang/cs685/schedule.html**

- **4/2:** Deadline of applying for the first round of API credit
  - https://piazza.com/class/m1kz66st9dn62i/post/146
- **4/9:** Midterm Review?
- **4/11**: HW 2 due
- **4/18 (**Friday but Monday Schedule**)**: **Midterm**
- **5/9: Final project report due**

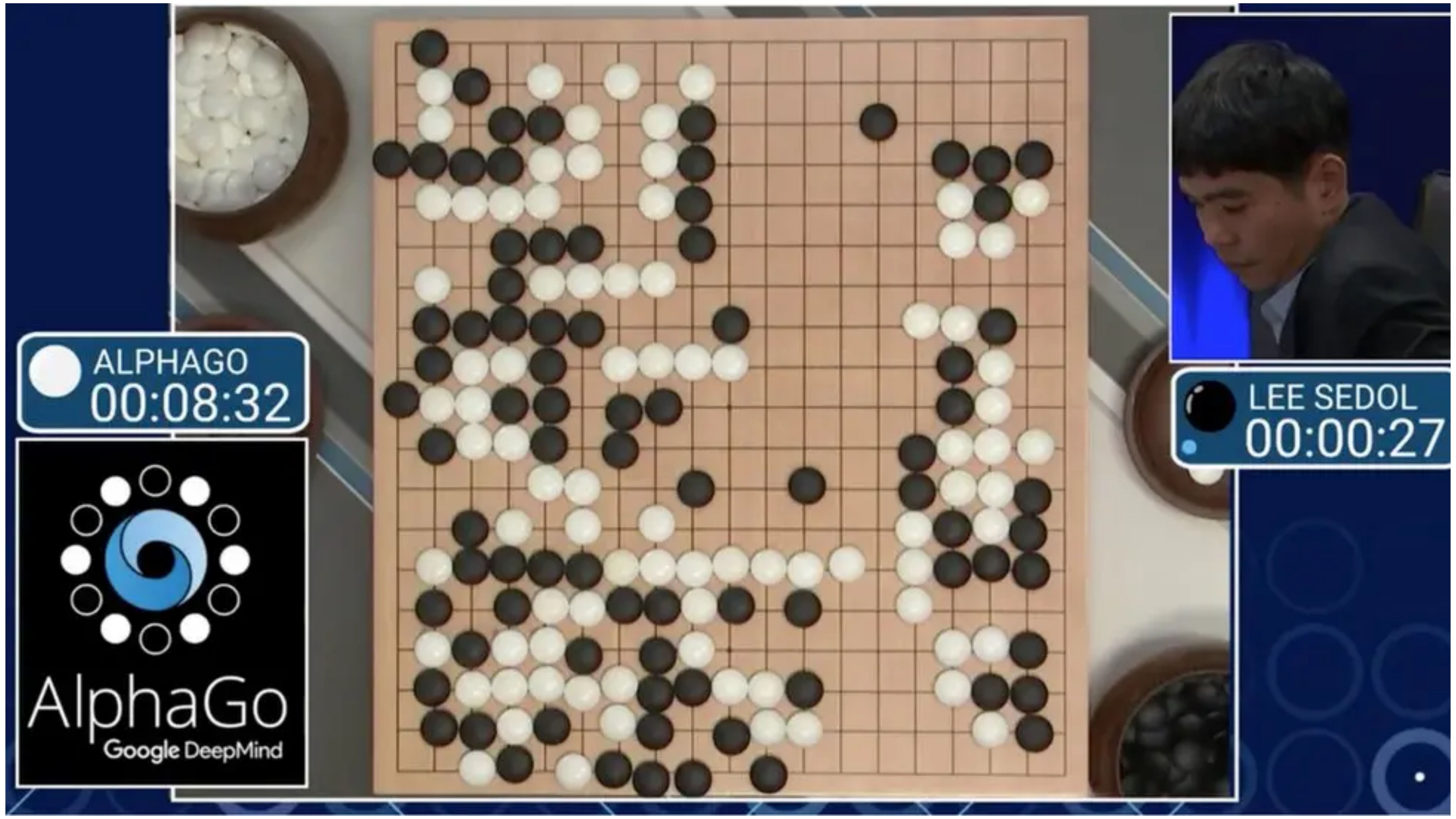
Difference between LLM and LRM

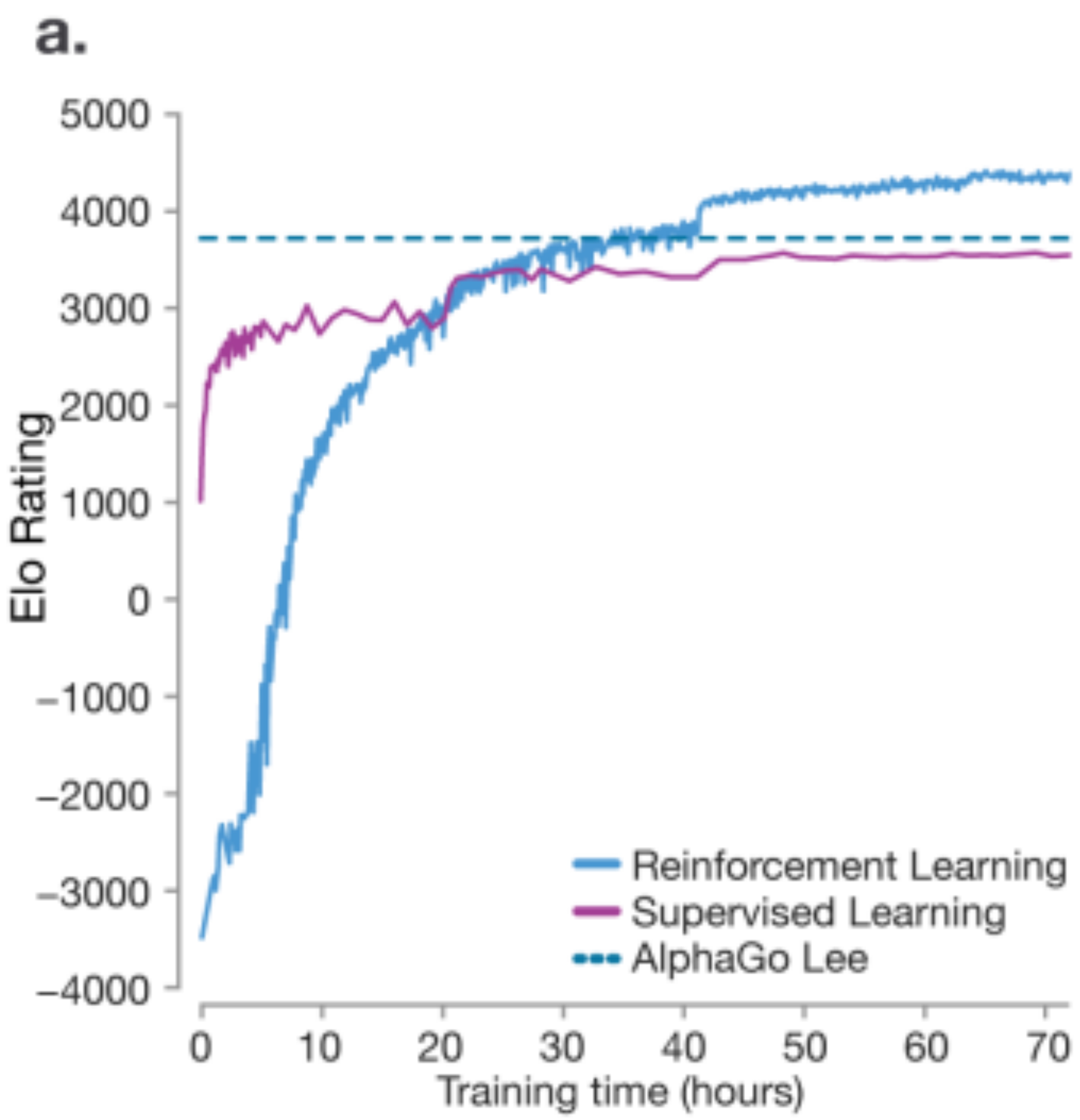
- Demo:
  - Please list all possible combinations of four different numbers from 1 to 10 (inclusive) such that their summation is 22.

https://chatgpt.com/

Do LRMs really Learn to Think like Humans?

AlphaGo and AlphaZero





Why can AlphaGo be better than top human players, but LRM cannot?

https://www.bbc.com/news/technology-35785875

https://www.science.org/doi/10.1126/science.aar6404?__cf_chl_tk=67Ig3VWHBOjaw3ybBhVGn2gbtd2QZ4UXUxDS21EBct4-1742742238-1.0.1.1-jx7XtvwAlV5eX51WMPAteOy04PT4tJF2e28qsLvXeTc

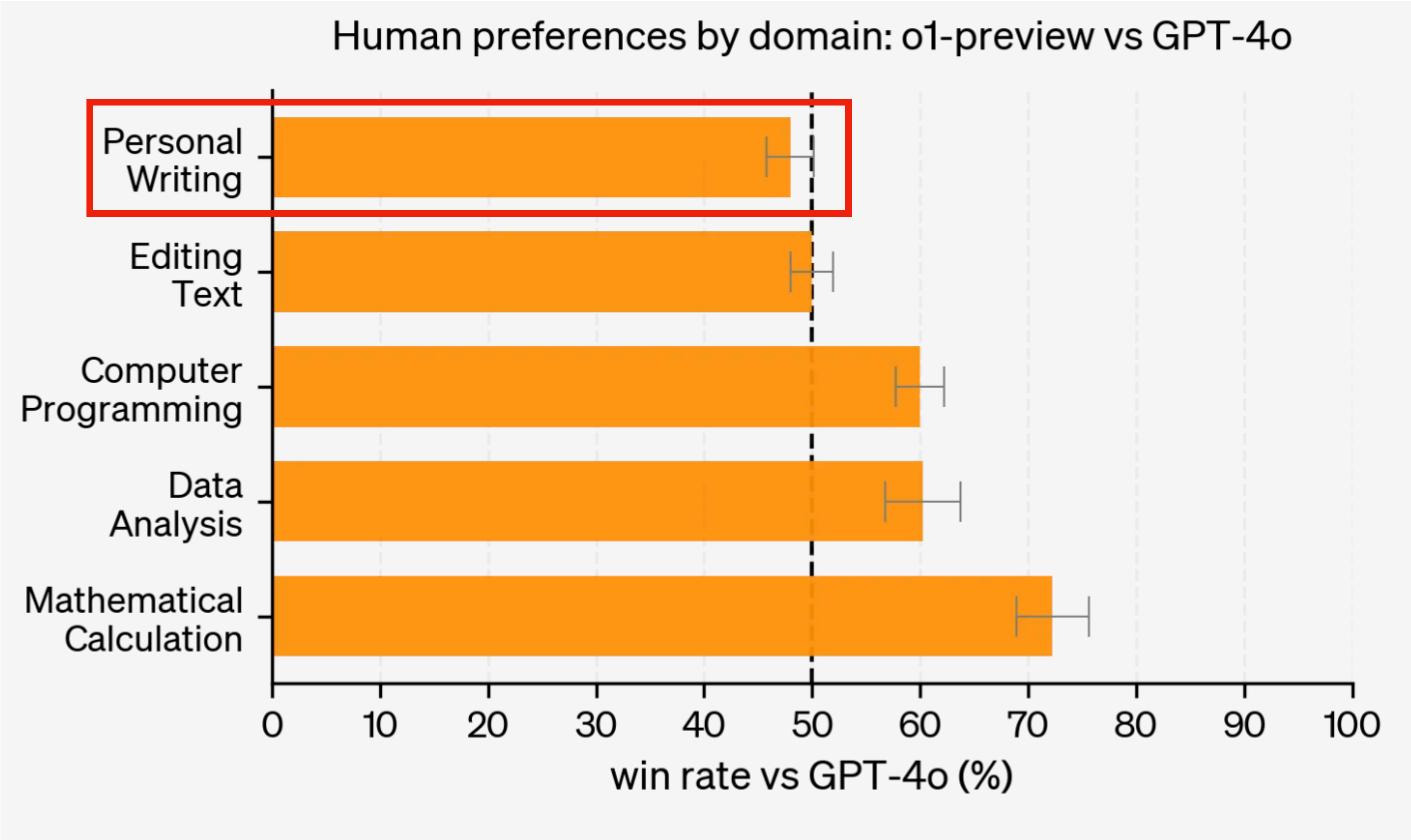
Evaluation Limitation

In many areas, we do not have evaluation functions that are cheap, reliable, and comprehensive

The Power of Evaluation Functions
(Will be Discussed More In the Future)

- Could be used in reinforcement learning
  - Math Answers, Winning of the Game, Reward Model for Alignment
- Could be used in best-of-N
- Could be used in evaluating the high-quality output
  - LLM as a judge for creative writing
- Could be used in evaluating the low-quality output

Better in Reasoning is not Better in Everything



	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

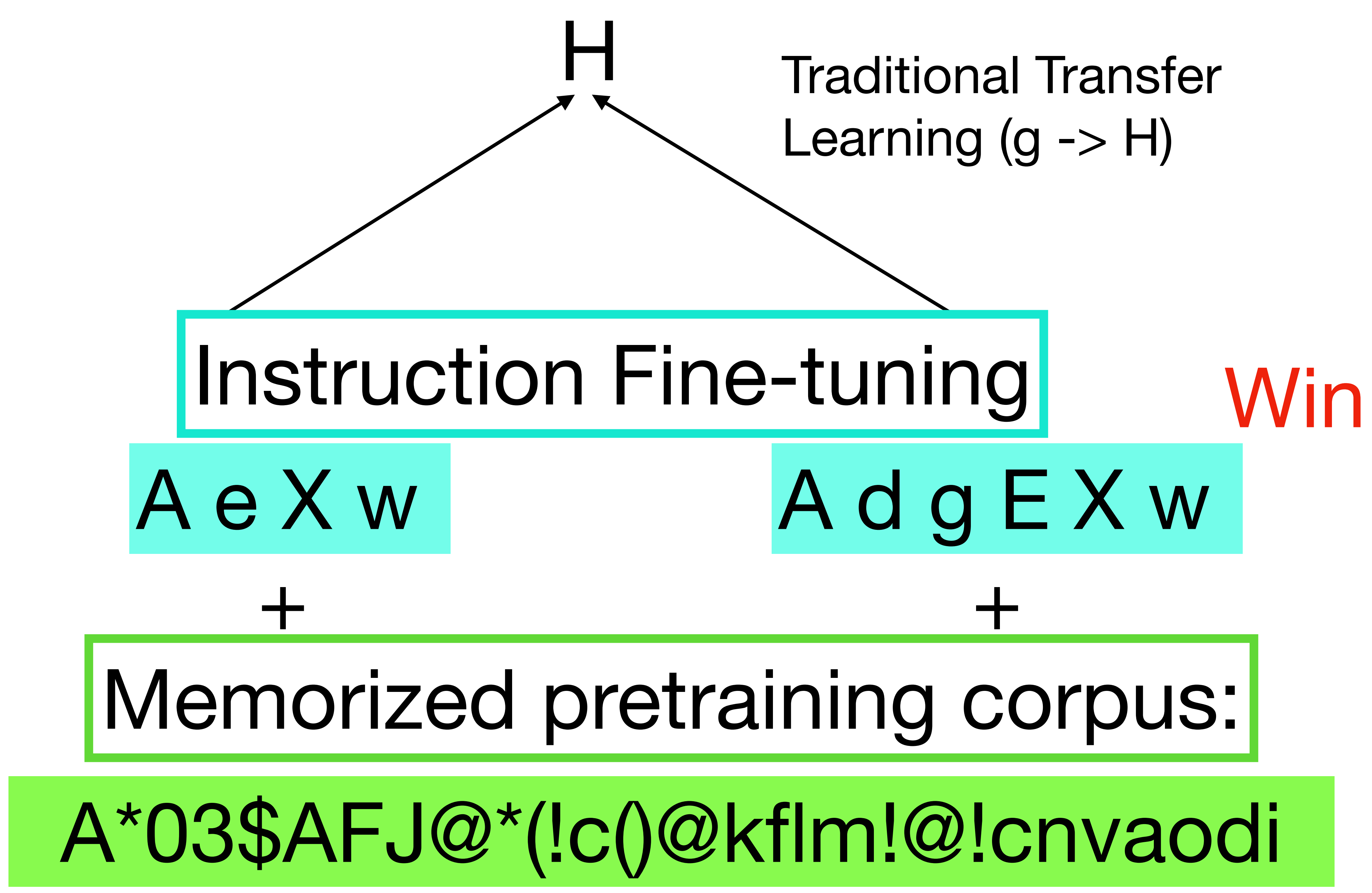
Table 4 | Comparison between DeepSeek-R1 and other representative models.

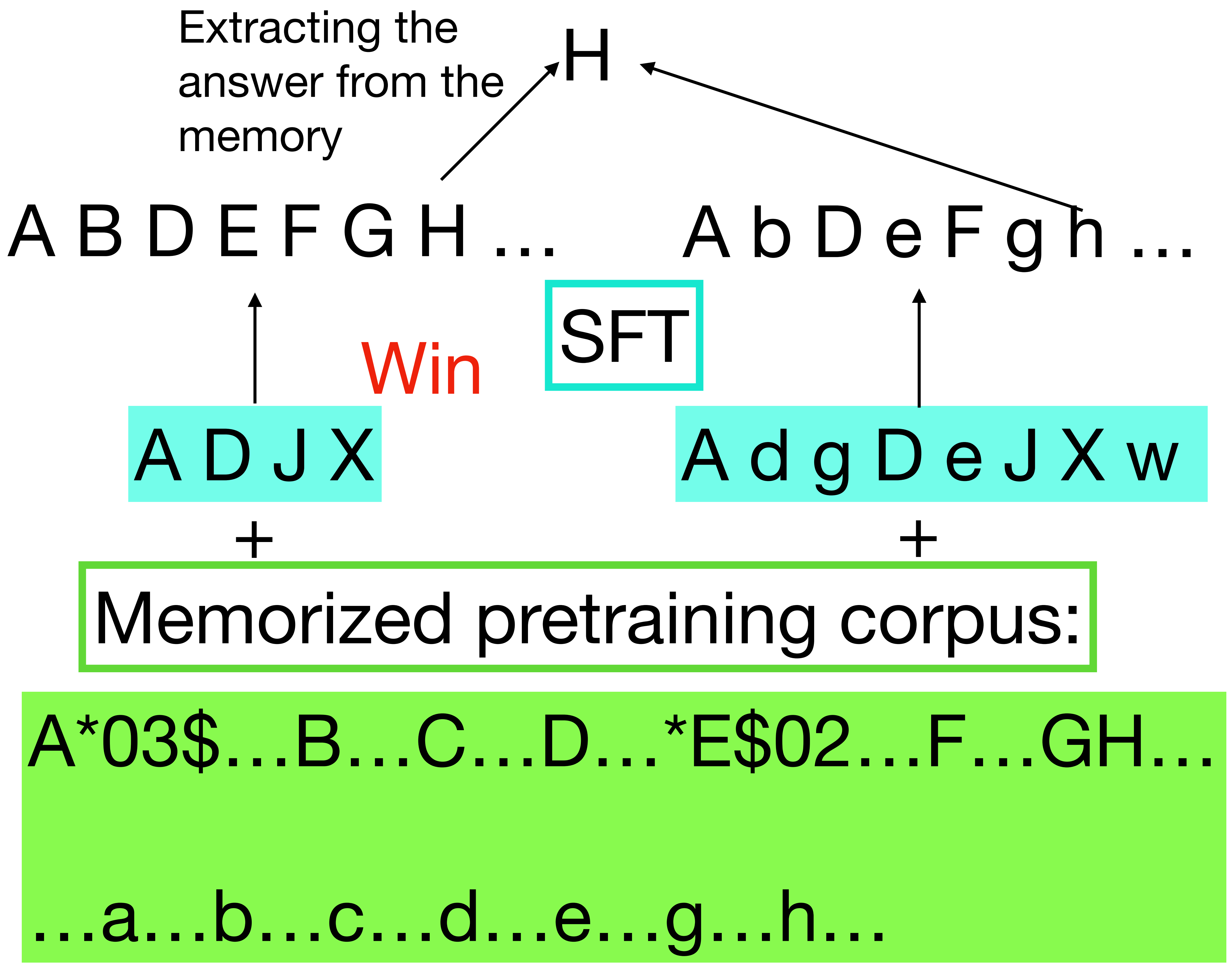
LLM Limitations

The reasoning performance still heavily depends on the pretraining, which suggests that LLM still struggles to generate something completely new

Why Could Fewer Data be Better?

- First task -> A: high-quality data, a: low-quality data





Recent studies show that such transfer learning does not actually work generally. See this paper:

Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning (https://arxiv.org/pdf/2305.11383)

Distillation

- First task -> A: high-quality data, a: low-quality data

🧑‍🎓 SLM

H

Instruction Fine-tuning

From LLM

+

Memorized pretraining corpus:

A*03$AFJ@*(!c()@kflm!@!cnvaodi

🧑‍🏫 LLM

H

A B D E F G H …

SFT

A D J X

+

Memorized pretraining corpus:

A*03$…B…C…D…*E$02…F…GH…

…a…b…c…d…e…g…h…

Deepseek R1 Distillation

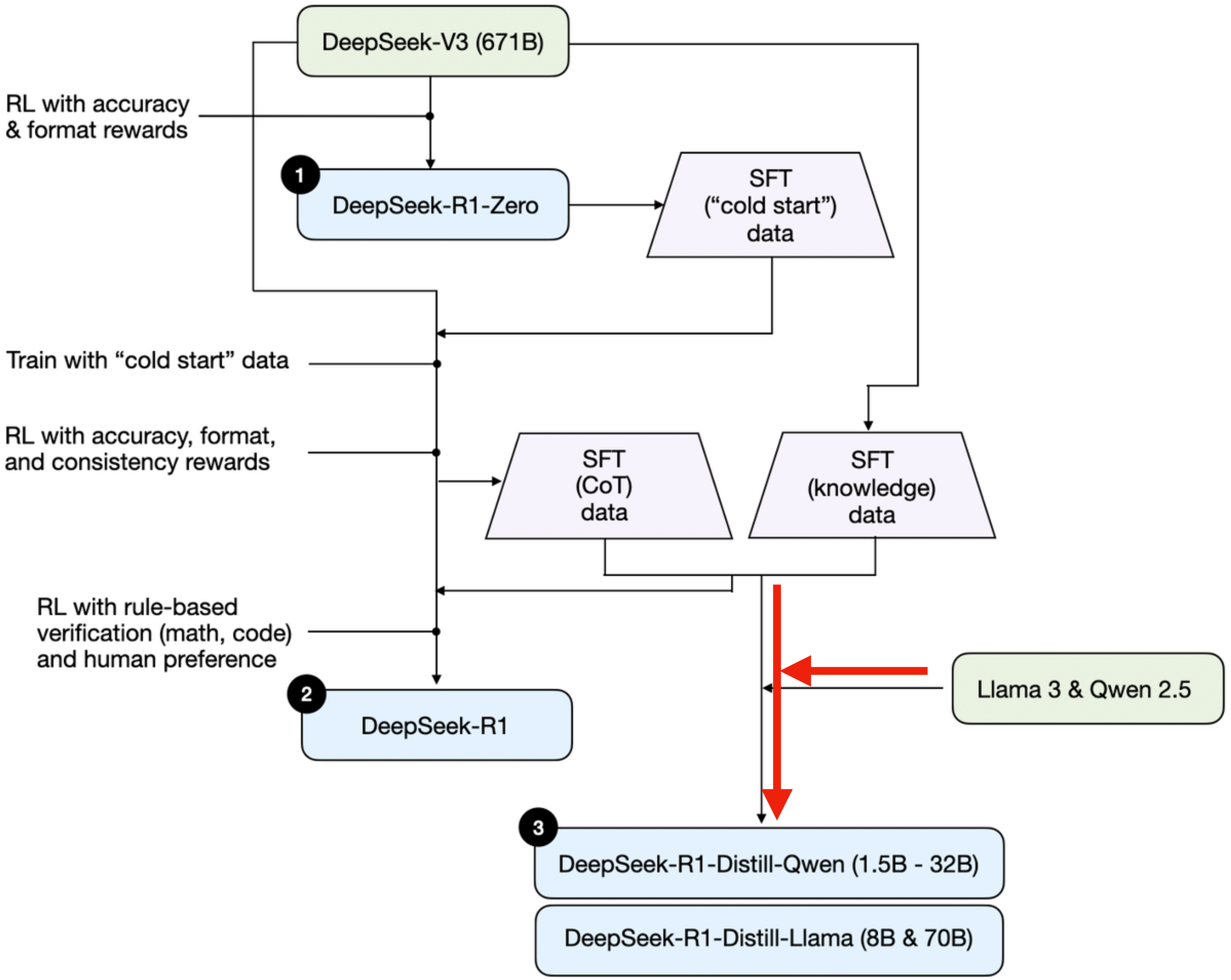
Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.

LLM size matters!

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.



Less is More for Distillation

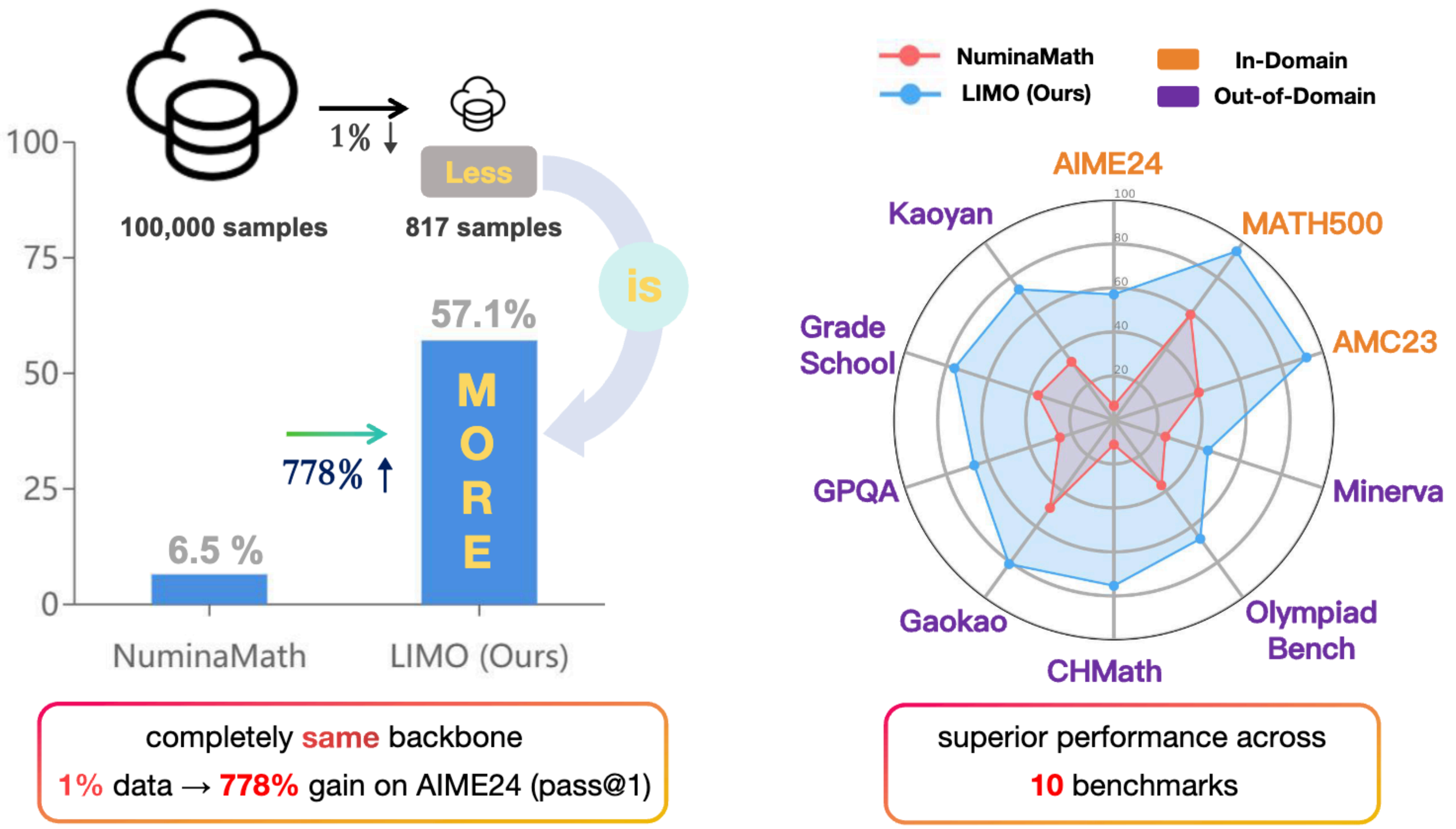


Figure 1: LIMO achieves substantial improvement over NuminaMath with fewer samples while excelling across diverse mathematical and multi-discipline benchmarks.

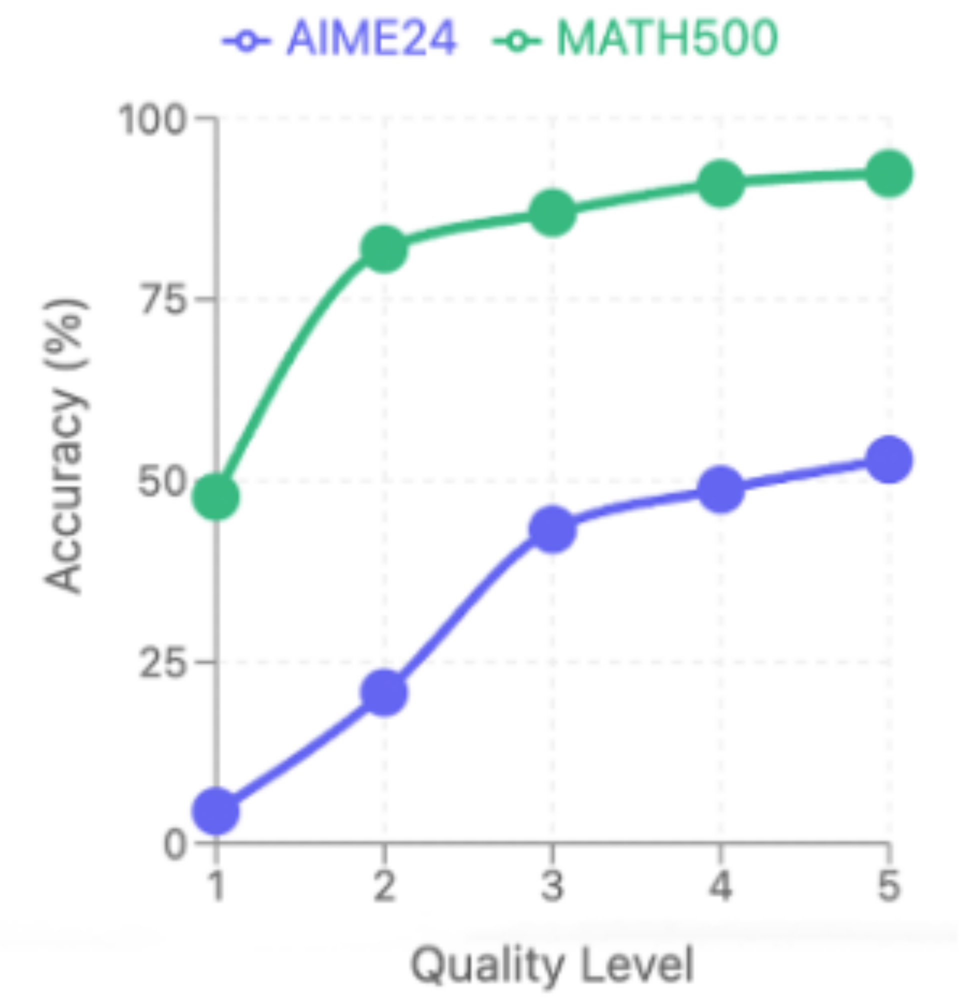


Figure 2: Comparison of models trained on reasoning chains of different quality levels.

LIMO: Less is More for Reasoning (https://arxiv.org/pdf/2502.03387)

Test-time Scaling from Distillation (s1)

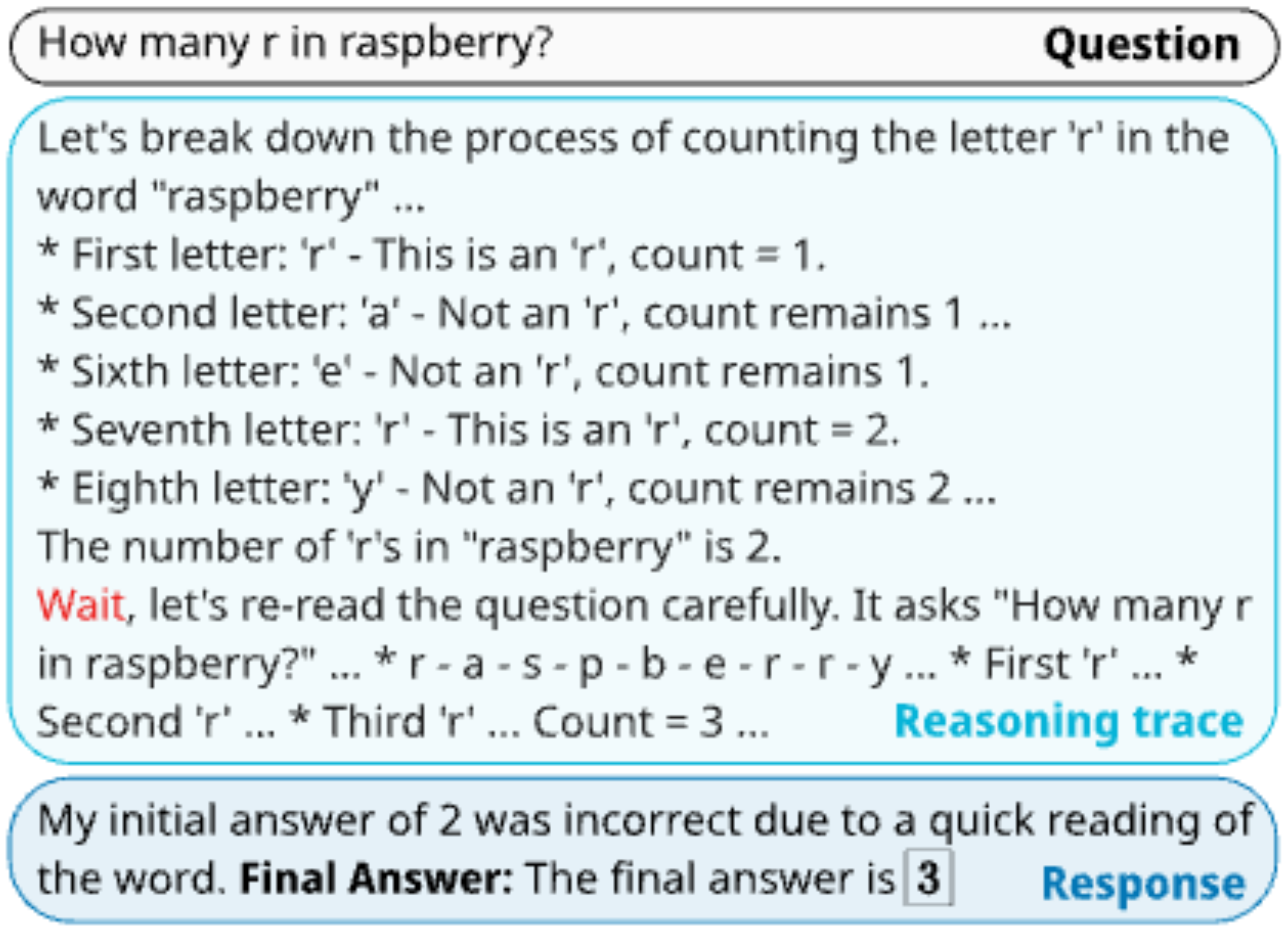
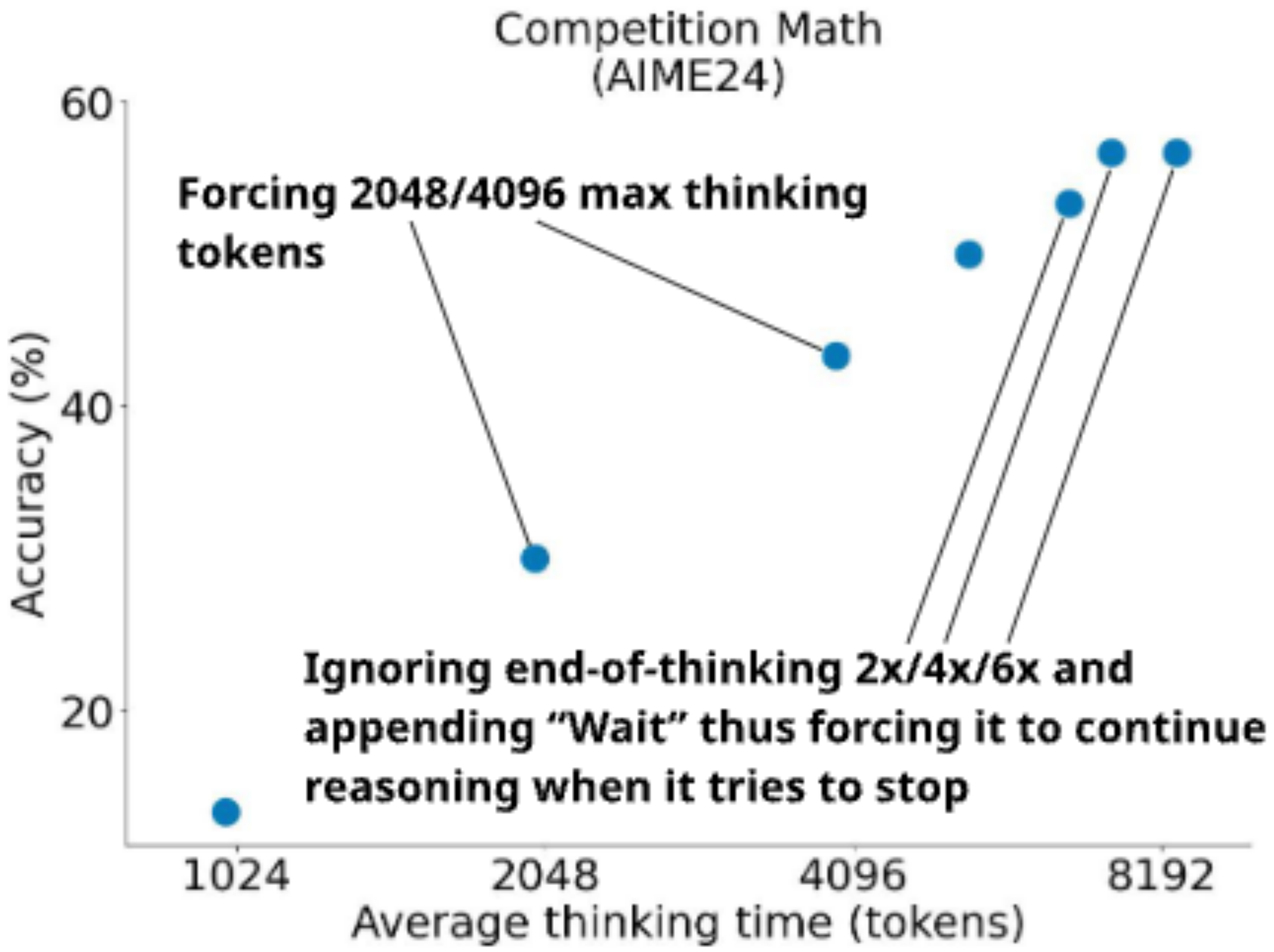


Figure 3. **Budget forcing with s1-32B.** The model tries to stop after "...is 2.", but we suppress the end-of-thinking token delimiter instead appending "Wait" leading **s1-32B** to self-correct its answer.



(a) Sequential scaling via budget forcing

s1: Simple test-time scaling (https://arxiv.org/pdf/2501.19393)

Distillation Data

- What are the good questions?
  - Quality
  - Difficulty
  - Diversity
- Where are the high-quality solutions from?
  - From Gemini 2.0 Flash Think
- More high quality data is still better

Table 1. **s1-32B is a strong open reasoning model.** We evaluate **s1-32B**, Qwen, and Gemini (some entries are unknown (N.A.), see §4). Other results are from the respective reports (Qwen et al., 2024; Team, 2024; OpenAI, 2024; DeepSeek-AI et al., 2025; Labs, 2025; Team, 2025). # ex. = number examples used for reasoning finetuning; BF = budget forcing. See §A for our better **s1.1** model.

Model	# ex.	AIME 2024	MATH 500	GPQA Diamond
API only				
o1-preview	N.A.	44.6	85.5	73.3
o1-mini	N.A.	70.0	90.0	60.0
o1	N.A.	**74.4**	**94.8**	**77.3**
Gemini 2.0 Flash Think.	N.A.	60.0	N.A.	N.A.
Open Weights				
Qwen2.5-32B-Instruct	N.A.	26.7	84.0	49.0
QwQ-32B	N.A.	50.0	90.6	54.5
r1	$\gg$800K	**79.8**	**97.3**	**71.5**
r1-distill	800K	72.6	94.3	62.1
Open Weights and Open Data				
Sky-T1	17K	43.3	82.4	56.8
Bespoke-32B	17K	**63.3**	93.0	58.1
s1 w/o BF	**1K**	50.0	92.6	56.6
s1-32B	**1K**	56.7	**93.0**	**59.6**

Private Data Seems to be Important

- We know the following facts
  - After RL for reasoning, every model achieves similar performances
    - Grok 3 / OpenAI o1 or o3 / Gemini 2.5 Pro / **Deepseek R1 / Qwen QwQ / Kimi**
    - Meta hasn’t had a reasoning model
  - OpenAI’s o1 sometimes outputs Chinese math solutions

Open Questions

- How do pretraining data or RL allow LLMs to generalize?
    - Are there very similar problems in the pretraining data and SFT just activates the memory of LLM or does reasoning ability emerge when the LLM becomes larger?
    - RL allows the LRM to come up with some novel reasoning paths?
    - Reasoning RL still mostly changes the style?
- Do LLMs piece the solutions of existing subproblems together using CoT?
- Is RL more generalizable than SFT? If yes, why?

Summary

- Without SFT, RL could discover very effective long CoT to solve the problems
    - RL could unlock some new reasoning abilities of LLMs
- Reasoning ability is not generalizable to other domains
    - LRM will still have some hallucinations and difficulty in following the constraints in other domains that do not have reliable evaluation functions.
- Using SFT in distillation, we only need very few high-quality data to learn to output such long CoT in a generalizable way
    - Lots of reasoning ability still depends on the pretraining stage