Incentivizing LLMs to Self-Verify Their Answers

Presenter: Fuxiang Zhang

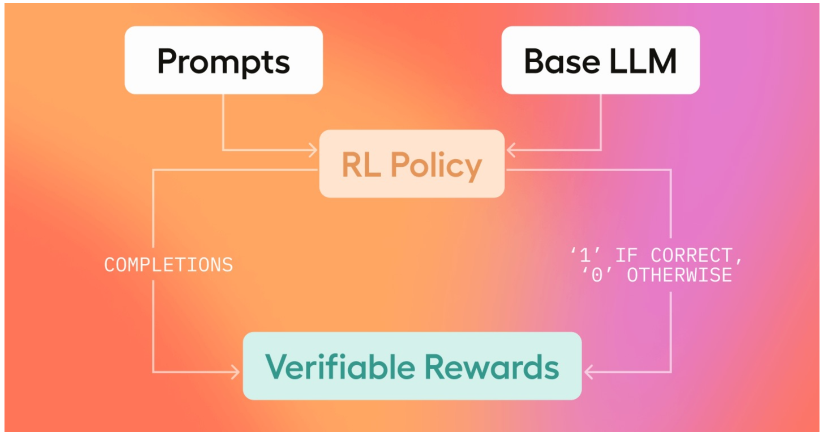
NeurIPS 2025

Fuxiang Zhang[1,2] **Jiacheng Xu**[1,2] **Chaojie Wang**[2] **Ce Cui**[2] **Yang Liu**[2] **Bo An**[1,2*]

[1] Nanyang Technological University, Singapore [2] Skywork AI

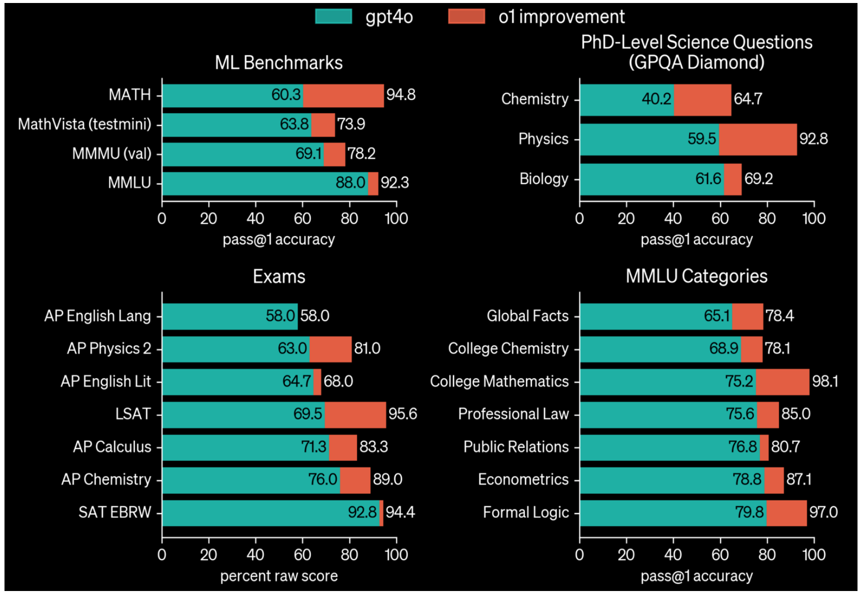
Large Language Models (LLM) for Reasoning

- LLMs are trained with Reinforcement Learning (RL)
  - Learned from the correctness of solutions (a *verifiable* reward)
  - Coding, mathematical problems, question-answering



RL from Verifiable Rewards

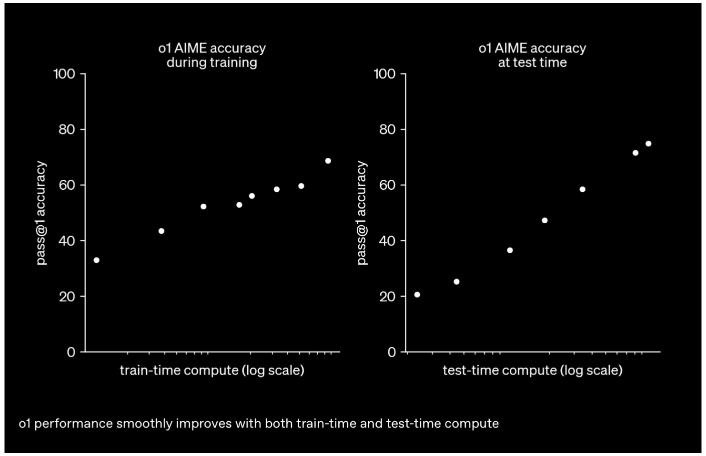
https://openai.com/index/learning-to-reason-with-llms



Reasoning model (o1) vs. non-reasoning model (gpt4o)

Test-time Scaling of Reasoning Models

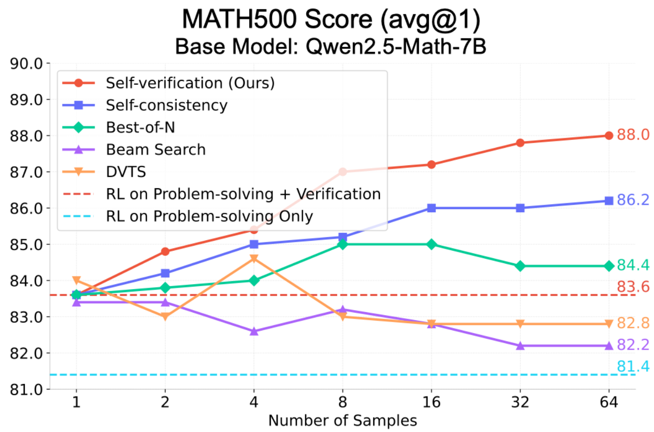
- Test-time improvement of LLMs:
  - Generate multiple answers and
    - Select the most consistent one
    - Use an external verifier to choose the answer with the highest score

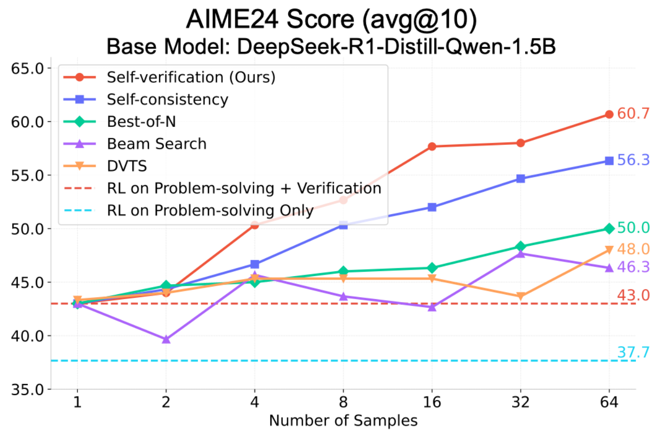


- Two paths of LLM scaling:
  - Training-time compute
  - Test-time compute
- Can we find a synergy of them?

Synergize Training- and Test-time Compute

- Can we find a synergy of training- and test-time compute?
  - Simple combination does NOT work!



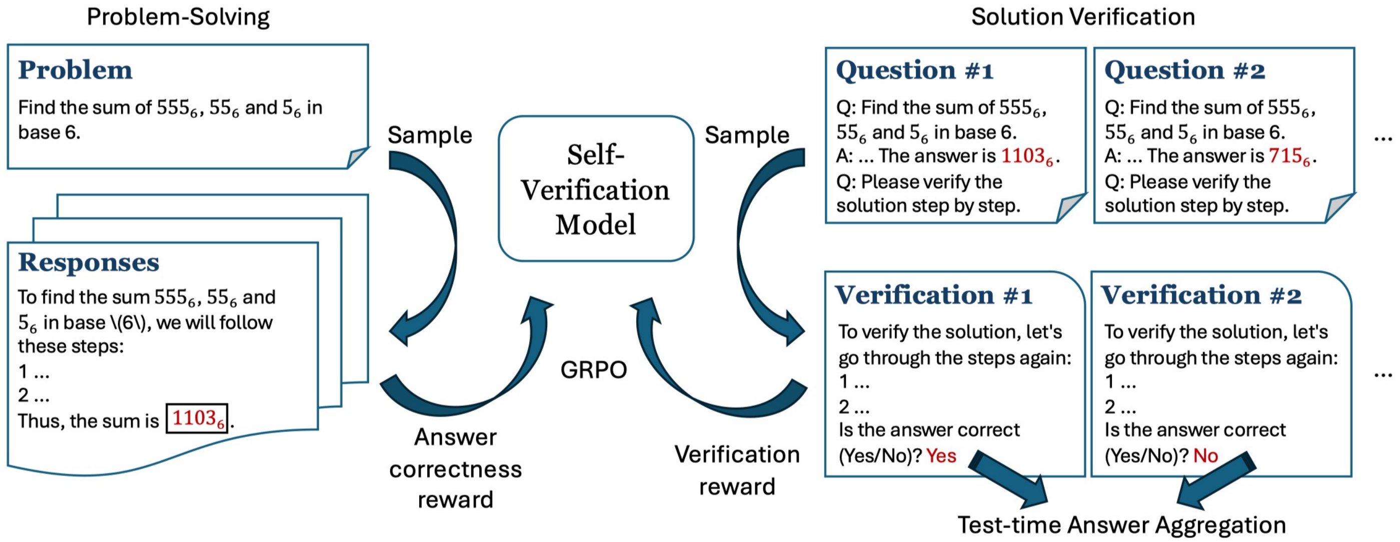


- External verifiers haven’t seen the solutions of RL models before (distribution shift)

Self-Verification Paradigm

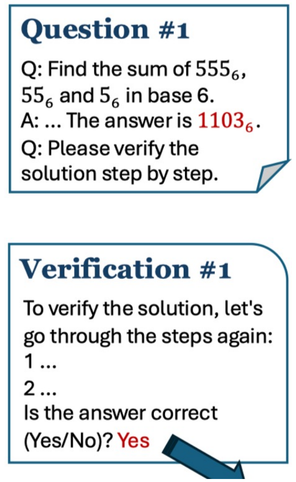
- How to deal with the distribution shift issue?
  - Train a unified model for both problem-solving and verification

The model itself knows how to verify its answer better.



Incentivizing the LLM to Verify Answers

- Generative Verification:
  - Ask the model to generate predicates and judgments
  - [thinking] Is the answer correct? (Yes/No)
- Keep the original capability as a reasoning model
  - A Unified RL process
  - Train problem-solving and verification simultaneously
    - Policy-aligned online data buffer
    - Dynamic verification reward



Experiment Results

Table 1: Average **greedy-decoding scores** of different models on math reasoning benchmarks after post-training. The best scores from each model series are highlighted in bold. For AIME24, AIME25, and AMC23, we report the average scores over 10 samples for each problem.

Model	MATH500	AIME24 (avg@10)	AIME25 (avg@10)	AMC23 (avg@10)	Olympiad Bench
Model Series: Qwen2.5-Math-7B					
Self-Verification-Qwen-7B (Ours) (Problem-solving + verification)	**83.60**	20.00	**16.67**	63.75	**34.81**
Qwen2.5-Math-7B (Base model)	62.00	14.67	5.00	45.25	17.63
GRPO-Qwen-7B (Problem-solving Only)	81.40	19.67	15.67	**65.50**	32.89
SimpleRL-Qwen-Math-7B ([28])	80.80	**23.33**	10.00	63.75	32.15
Model Series: DeepSeek-R1-Distill-Qwen-1.5B					
Self-Verification-R1-1.5B (Ours) (Problem-solving + verification)	**87.00**	**43.00**	**31.33**	**77.50**	**44.30**
R1-Distill-Qwen-1.5B (Base model)	80.00	24.33	25.00	64.25	32.89
GRPO-R1-1.5B (Problem-solving only)	**87.00**	37.67	26.67	72.50	40.74
DeepScaleR-1.5B-Preview ([29])	83.00	37.00	31.00	77.25	43.56

Better greedy-decoding scores after RL!

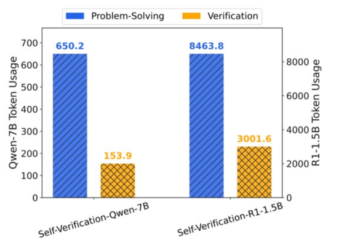


Figure 3: Token usage comparison between problem-solving and verification tasks.

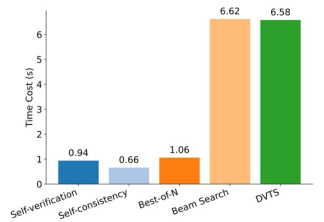


Figure 4: Average time cost of different test-time scaling methods per problem from MATH500.

Efficient token usage and time cost!

Table 2: The performance of different models on verifying MATH500 solutions generated by the Self-Verification-Qwen-7B model. We highlight the best scores from the open-source models in bold.

Category	**Method**	**Accuracy**	**F1 Score**
Open-source Models (~7B)	Self-Verification-Qwen-7B (Ours)	**87.20**	**92.83**
	Qwen2.5-Math-7B (Base model)	73.20	84.93
	Llama-3.1-8B-Instruct	67.00	78.20
Proprietary Models	GPT-4o	85.20	91.57
	Claude-3.7-Sonnet	90.20	94.46
	DeepSeek-v3	89.00	93.73

Table 3: The performance of different models on verifying AIME24 solutions generated by the Self-Verification-R1-1.5B model. We highlight the best scores from the open-source models in bold.

Category	**Method**	**Accuracy**	**F1 Score**
Open-source Models (1.5B or ~7B)	Self-Verification-R1-1.5B (Ours)	**56.67**	**67.72**
	R1-Distill-Qwen-1.5B (Base model)	38.00	49.46
	R1-Distill-Qwen-7B	46.00	59.50
	Llama-3.1-8B-Instruct	55.67	45.71
Proprietary Models	GPT-4o	59.33	65.54
	Claude-3.7-Sonnet	64.33	71.16
	DeepSeek-v3	57.67	66.67

Better verification accuracy even than GPT-4o!

Table 4: Average **test-time scaling** scores of different methods on various math reasoning benchmarks. All the test-time scaling methods have a budget of 16 samples for each problem. The best scores from each model series are highlighted in bold. For AIME24, AIME25, and AMC23, we report the average scores over 10 samples for each problem.

Method@16	MATH500	AIME24 (avg@10)	AIME25 (avg@10)	AMC23 (avg@10)	Olympiad Bench
Model Series: Qwen2.5-Math-7B					
Self-Verification (Ours)	**87.20**	**26.67**	19.00	**73.25**	**39.70**
Self-Consistency	86.00	23.67	**21.67**	71.25	39.11
Best-of-N	85.00	23.33	16.67	64.25	37.92
Beam Search	82.80	21.00	15.00	67.25	35.71
DVTS	82.80	21.67	20.33	67.50	35.85
Model Series: DeepSeek-R1-Distill-Qwen-1.5B					
Self-Verification (Ours)	**93.60**	**57.67**	**37.67**	**92.00**	**50.96**
Self-Consistency	91.00	52.00	36.67	87.50	47.56
Best-of-N	86.00	46.33	33.67	83.75	43.41
Beam Search	88.40	42.67	33.00	85.25	44.59
DVTS	90.80	45.33	32.33	82.50	45.18

Better test-time performance using 16 generations!