清華大學
Tsinghua University

Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?

Yang Yue*, Zhiqi Chen*, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang✉

NEURAL INFORMATION PROCESSING SYSTEMS

Best Paper Runner-up Award

Presenter: Yang Yue

Tsinghua University

Background: LLM Reasoning

OpenAI o1

Deepseek-R1

Gemini Thinking

IMO Medal

ICPC Programming Medal

Background: Reinforcement Learning with Verifiable Reward

- The key to the success of reasoning models: **Reinforcement Learning with Verifiable Reward (RLVR)**

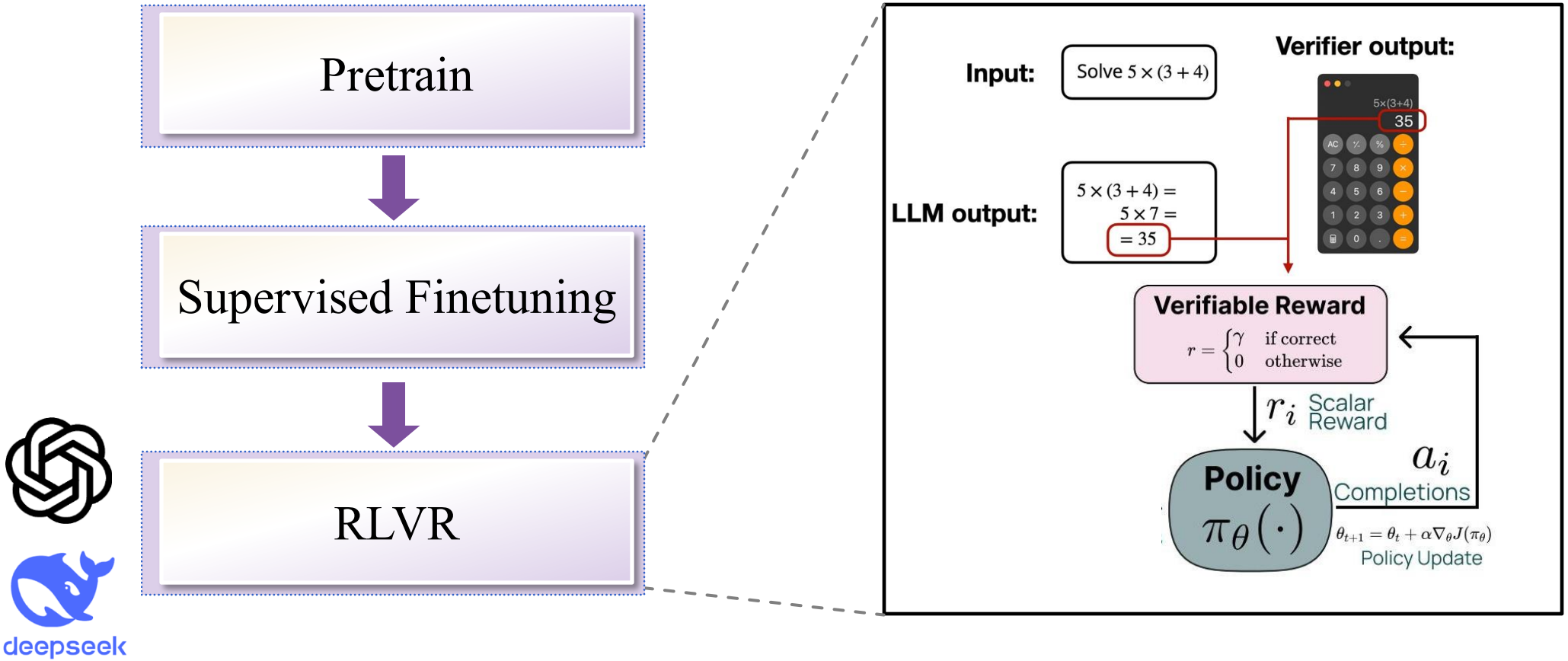
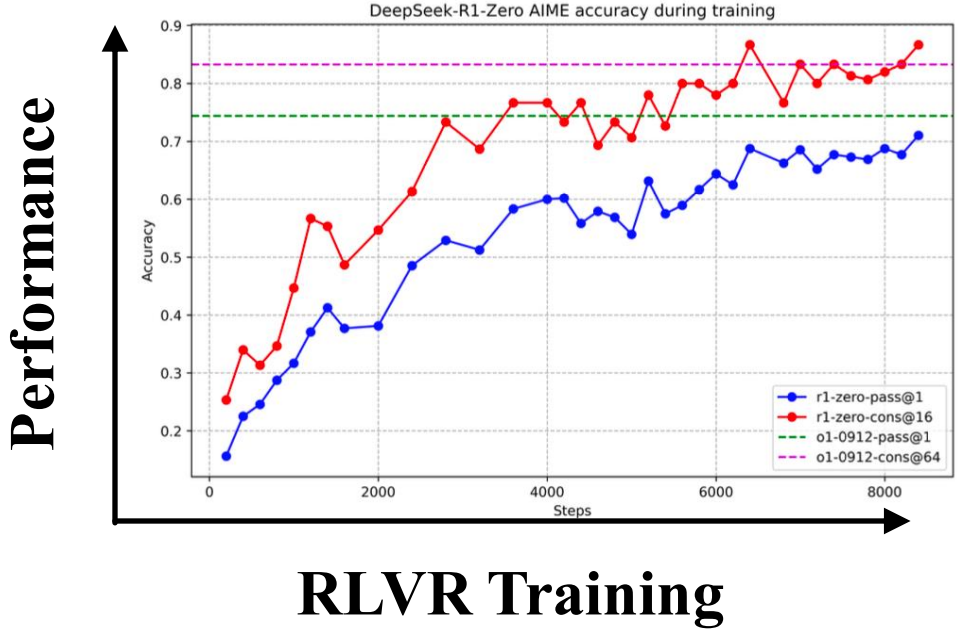


Image credit: https://magazine.sebastianraschka.com/p/the-state-of-llm-reasoning-model-training

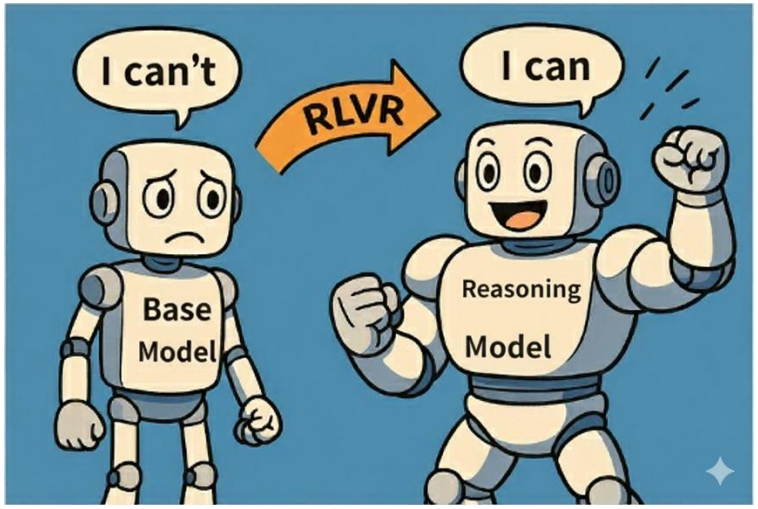
Lambert, Nathan, et al. "Tulu 3: Pushing frontiers in open language model post-training." *arXiv:2411.15124* (2024).

Background: Reinforcement Learning with Verifiable Reward

- **Reinforcement Learning with Verifiable Reward (RLVR)**
  - Self-generated data; no need for human-annotated CoTs
  - Significantly boost reasoning performance



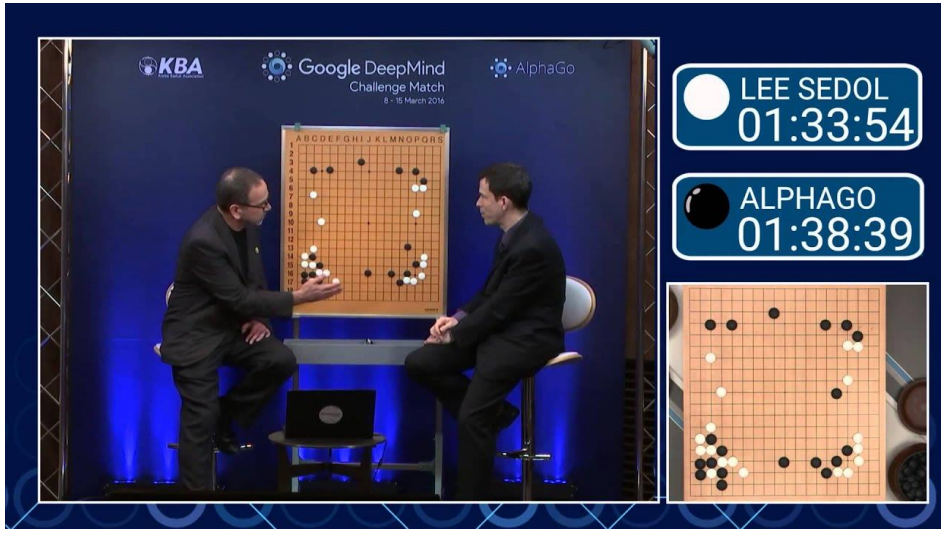
RLVR Training



Question

DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

AlphaGo's 37th move - a groundbreaking strategy that the agent discovered on its own



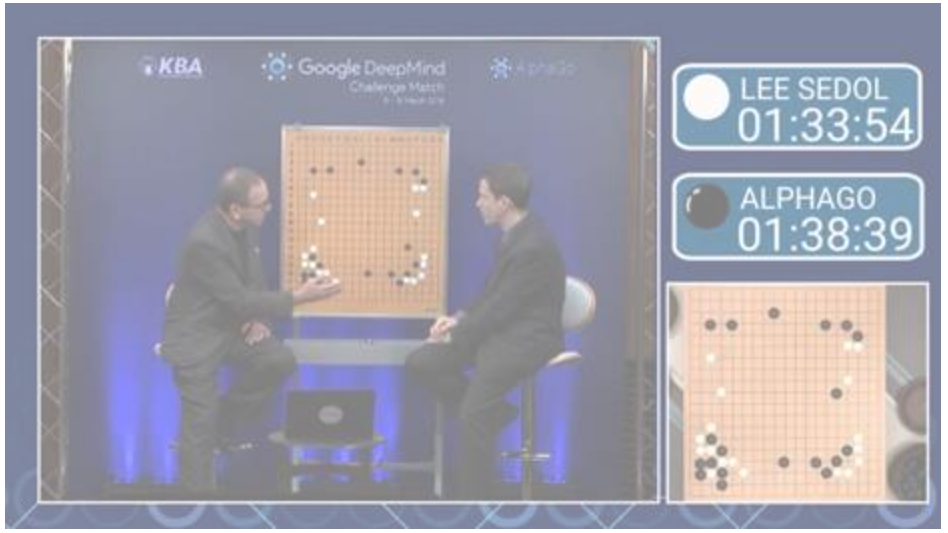
Question

Have we reached the *"AlphaGo Moment" for LLMs* yet?

Does RLVR truly discover new reasoning paths beyond the base model?

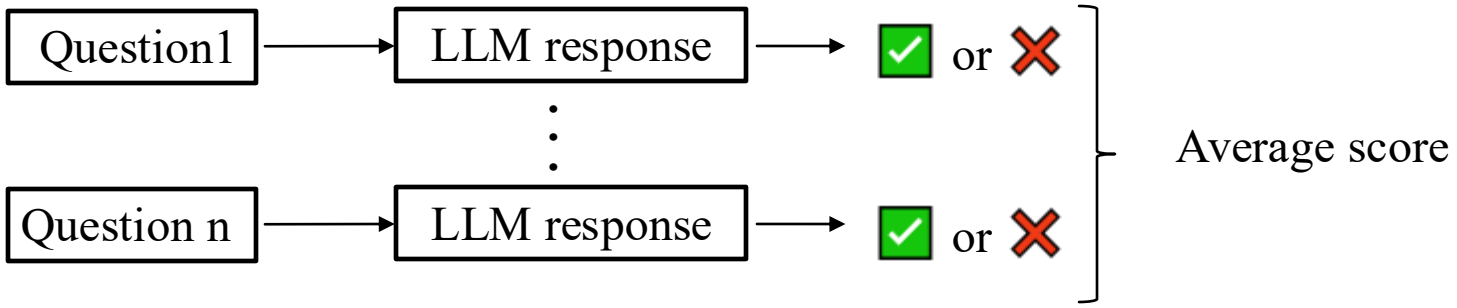
DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning

AlphaGo's 37th - a groundbreaking strategy that the agent discovered on its own



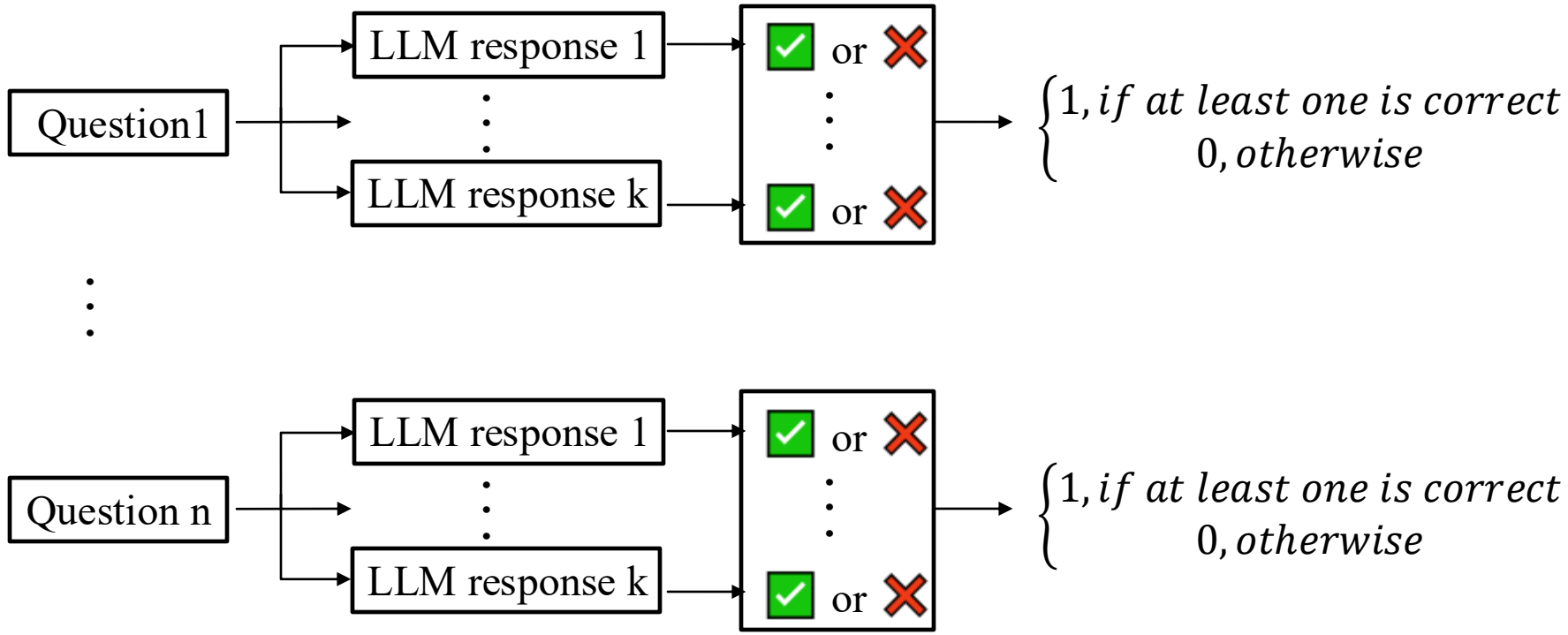
Metrics for LLM Reasoning Capacity Boundary

- Traditional metric for model reasoning performance: **Avg@k measures average capability**



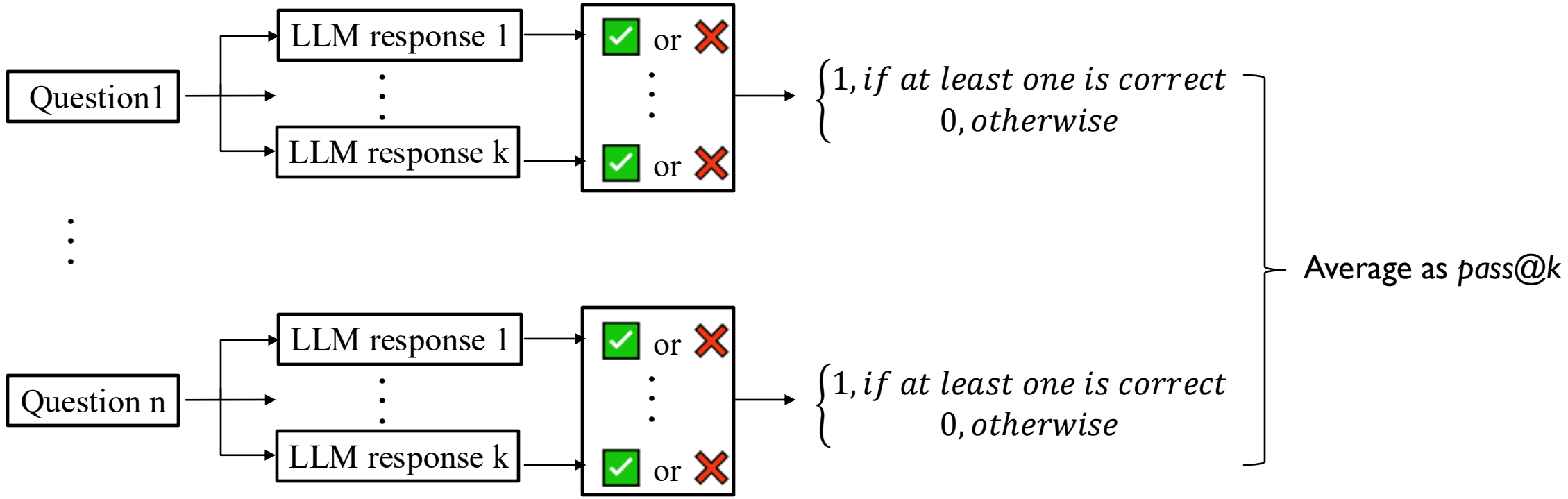
Metrics for LLM Reasoning Capacity Boundary

- Traditional metric for model performance: **Avg@k measures average capability**
- We use **pass@k** to **measure the capability boundary of models**



Metrics for LLM Reasoning Capacity Boundary

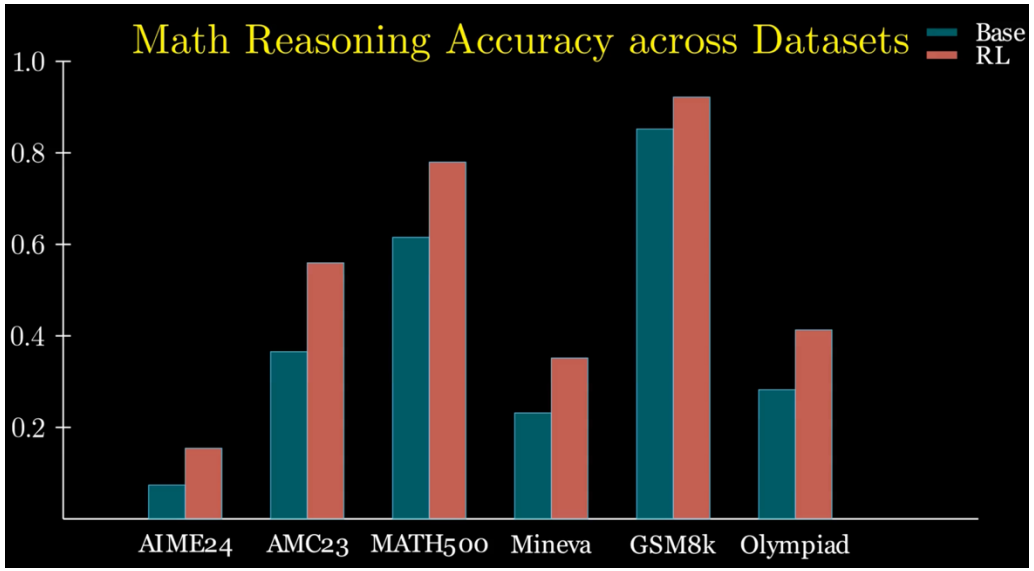
- Traditional metric for model performance: **Avg@k measures average capability**
- We use **pass@k to measure the capability boundary of models**



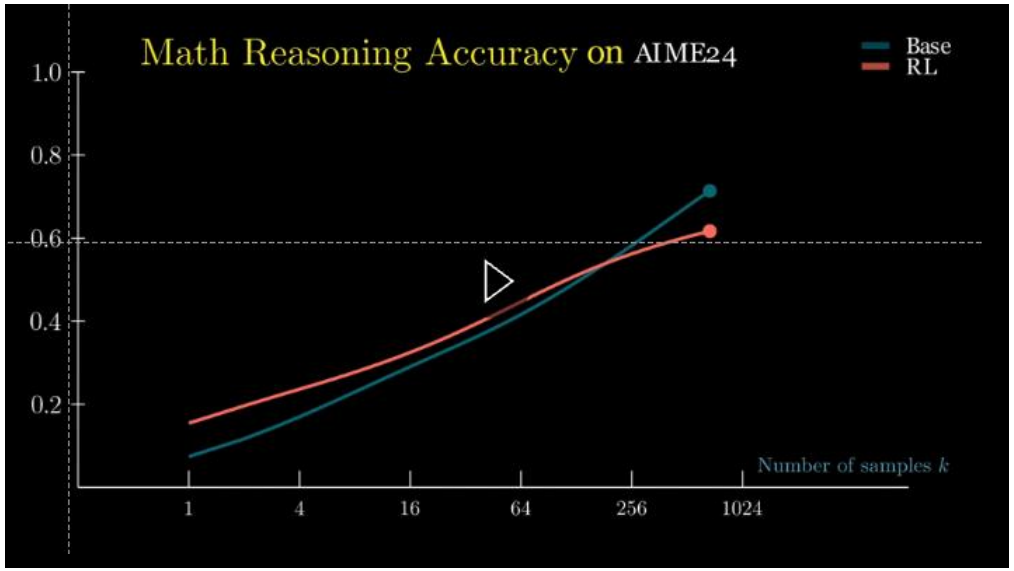
pass@k represents the proportion of problems in a dataset that the model can solve within *k* attempts.

Metrics for LLM Reasoning Capacity Boundary

- Avg@k measures average capability



- pass@k measures the capability boundary

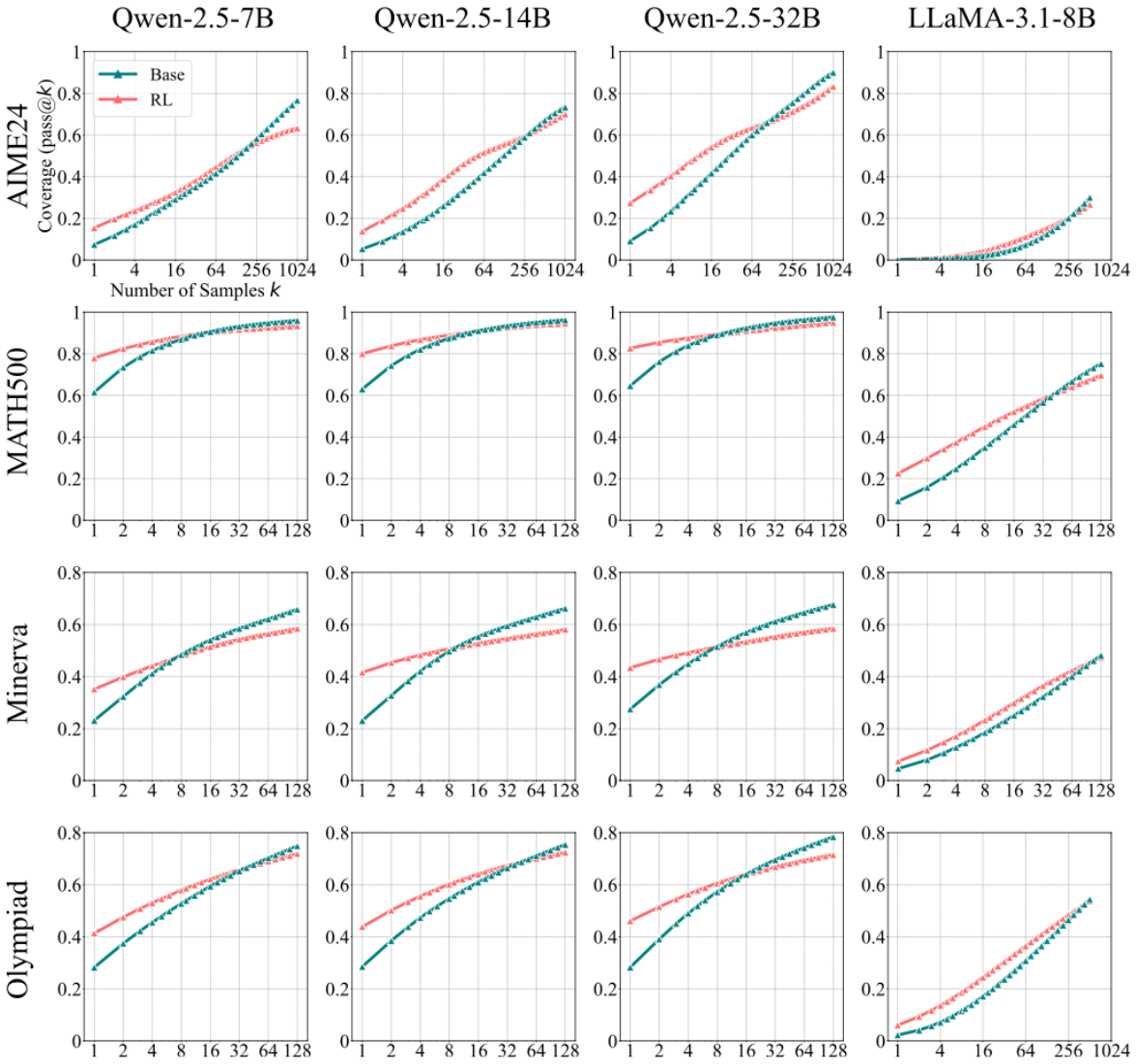


Surprisingly, the RLVR model underperforms the base model as k increases

RLVR's Effect on Reasoning Capacity Boundary

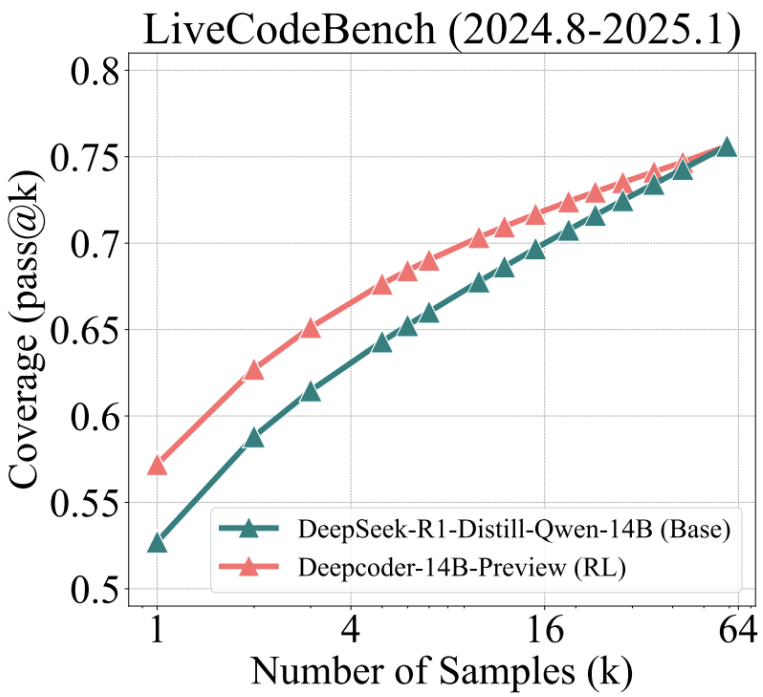
- **Base models *vs.* RLVR models**
  - Families (Qwen, LLaMA, Mistral, …)
  - Scales (7B, 14B, 32B, 72B, …)
  - Algorithms (PPO, GRPO, Reinforce++, …)
  - Domains (math, code, visual reasoning)
  - Benchmarks (AIME, MATH500, Minerva, …)

- Consistently observed that
  - For large k, pass@k of RL models does ***not*** surpasses base models
  - The number of solvable problems does ***not*** **increases after RLVR training**.

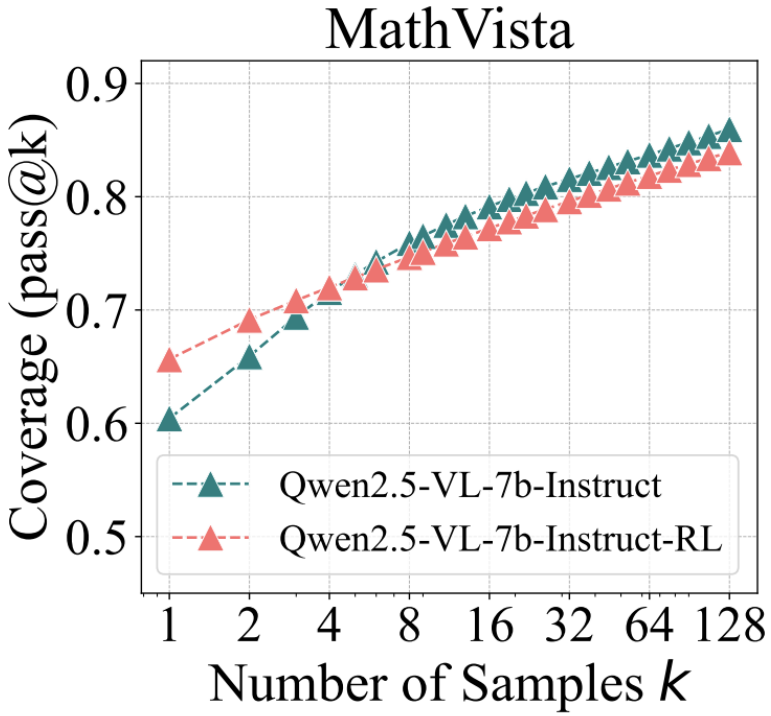


RLVR's Effect on Reasoning Capacity Boundary

- **Coding task**

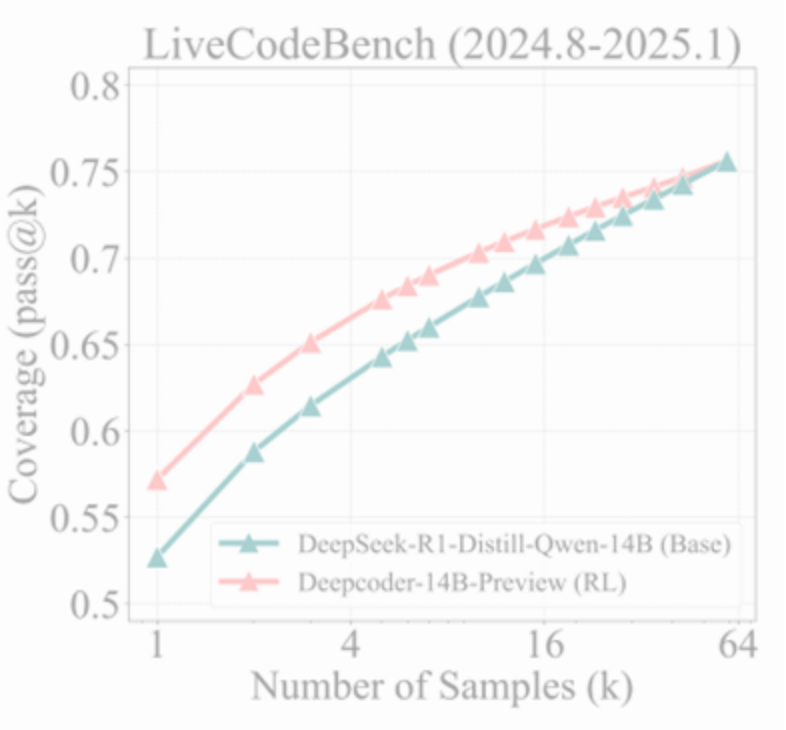


- **Visual reasoning task**

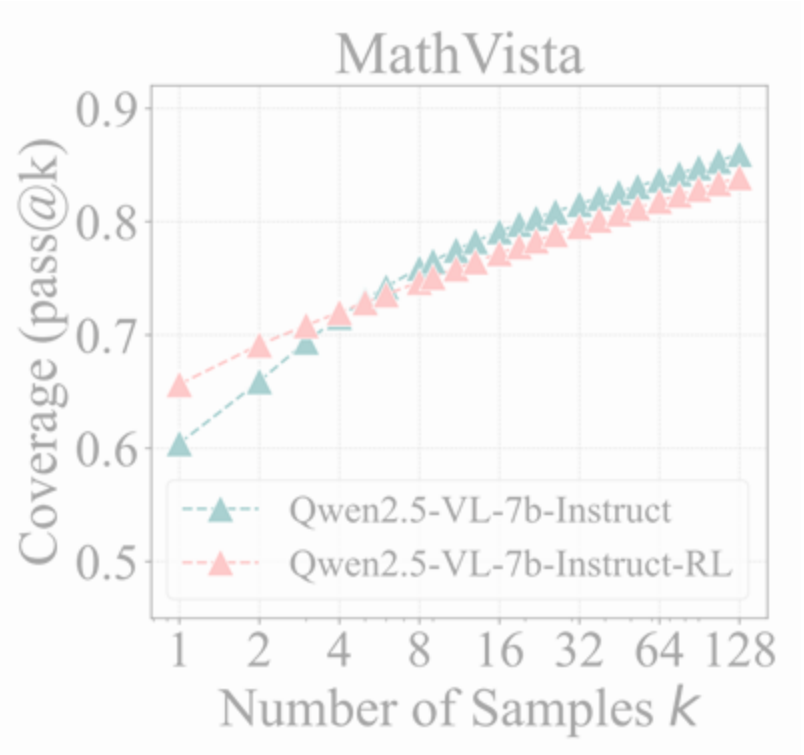


RLVR's Effect on Reasoning Capacity Boundary

☐ Coding task



☐ Visual reasoning task

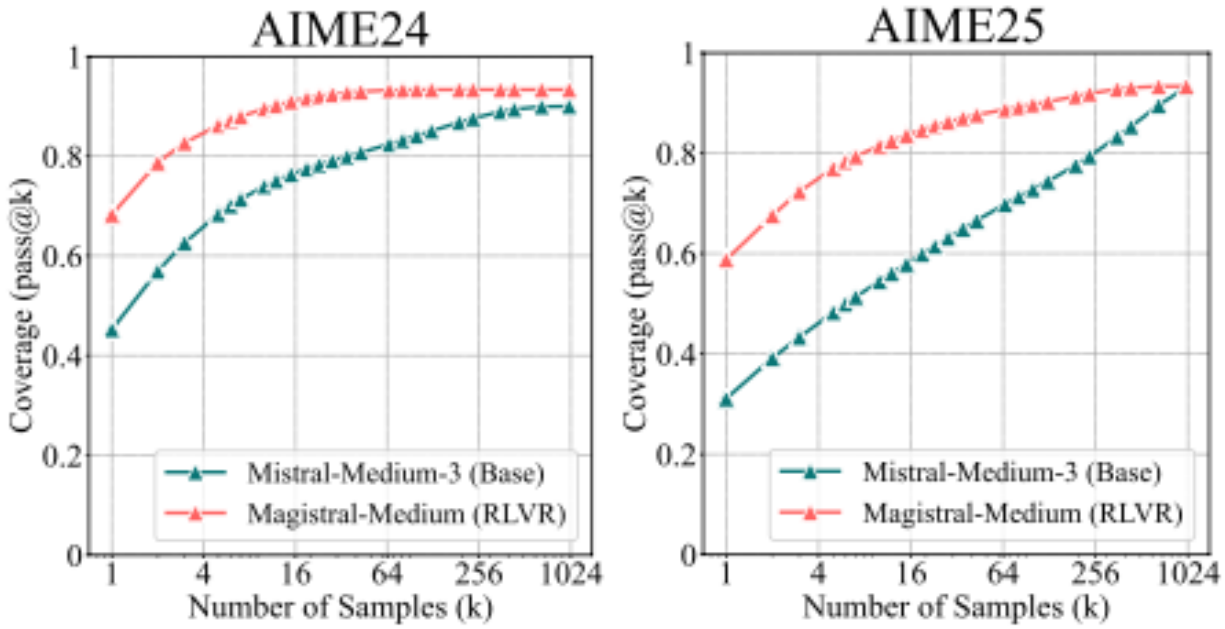


We try our best to rule out random guessing issue by:

1. Coding tasks
   - ➢ pass all unit tests
2. Datasets such as Minerva
   - ➢ answers with complex forms
3. Manually check a subset of CoTs for AIME

A preliminary experiment on model size scaling

- Magistral-medium (25.06)
  - A pure RLVR model
  - near-frontier performance in reasoning



The conclusion currently holds for highly capable, near-frontier reasoning models

Solvable problem proportion analysis

- **There exist some problems that base model can solve but RL model can't**

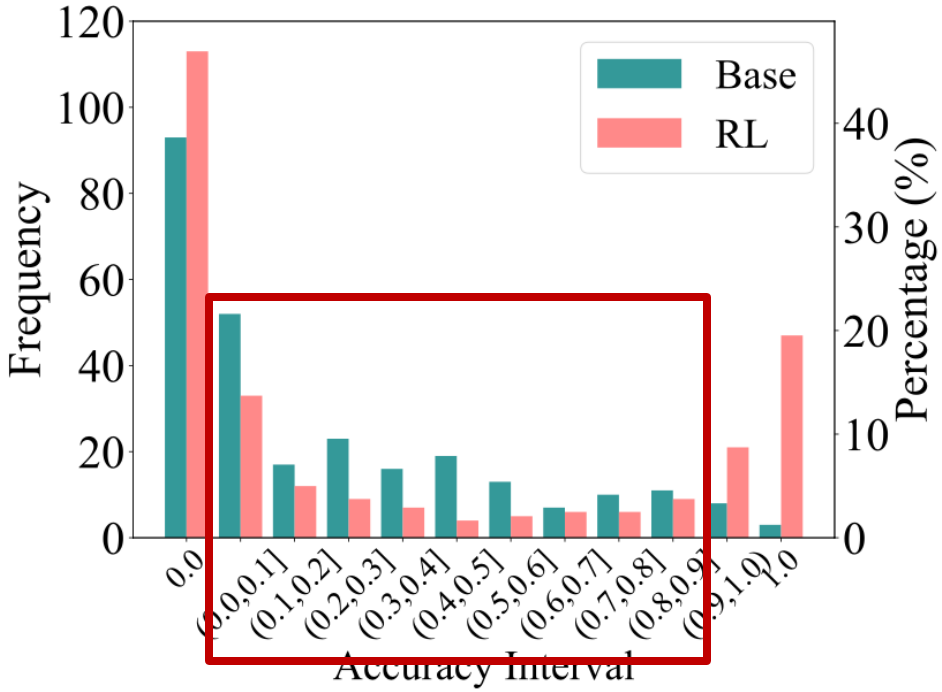
Base	SimpleRLZoo	AIME24	MATH500
✓	✓	63.3%	92.4%
✓	✗	13.3%	3.6%
✗	✓	0.0%	1.0%
✗	✗	23.3%	3.0%

Solvable problem proportion analysis

- There exist many problems that base model can solve but RL model can't
- **There are very few problems that RL model can solve but base model can't**

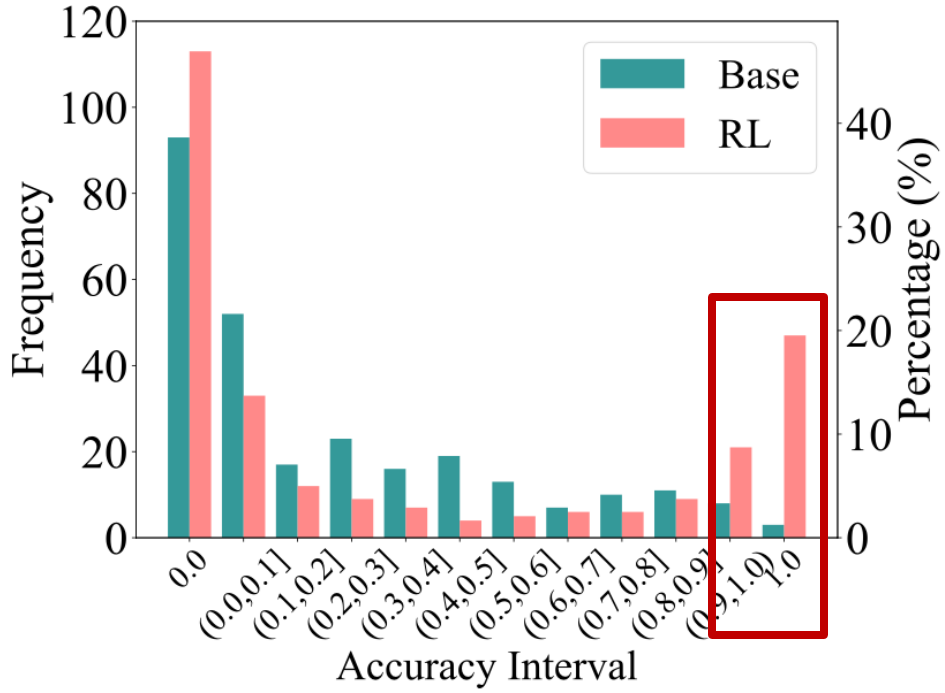
Base	SimpleRLZoo	AIME24	MATH500
✓	✓	63.3%	92.4%
✓	✗	13.3%	3.6%
✗	✓	0.0%	1.0%
✗	✗	23.3%	3.0%

Accuracy distribution analysis



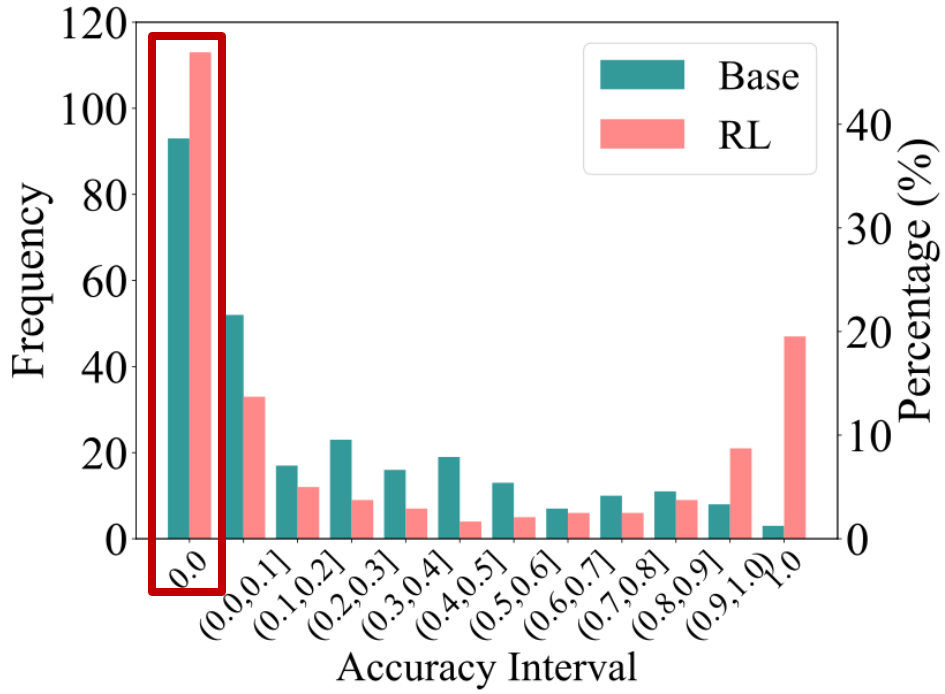
There exist many problems that base models can already solve, but only with low success rates (0 < accuracy < 0.9).

Accuracy distribution analysis



RLVR improves sample efficiency on these problems, raising accuracy to the 0.9–1.0 range.

Accuracy distribution analysis



However, RL model failed on the problems where base model has accuracy 0

Current RLVR gains mainly come from improved sampling efficiency, and rarely expand reasoning boundary.

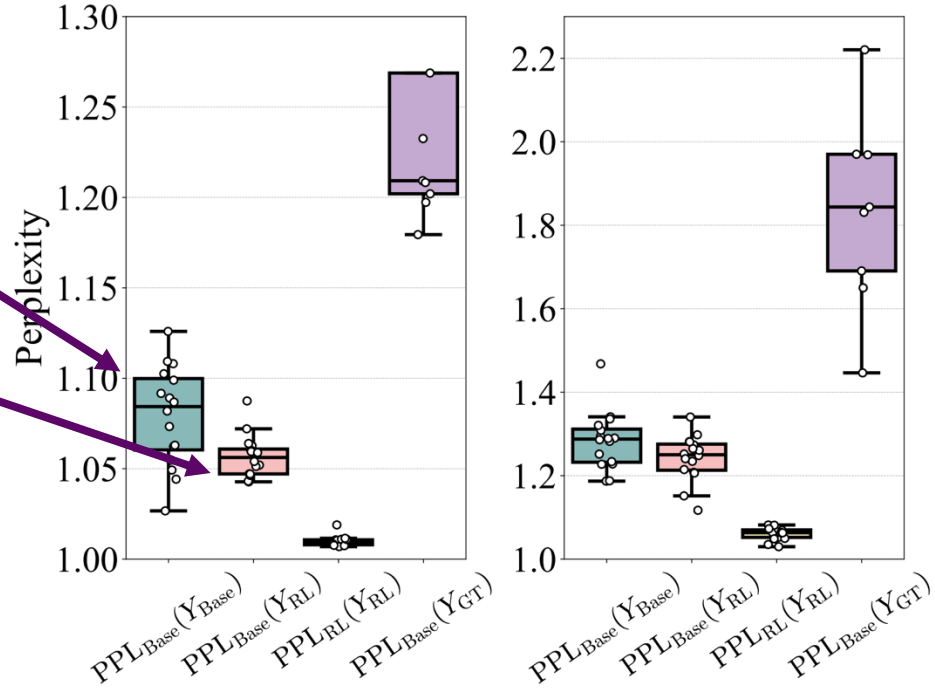
Perplexity analysis

- **Observation: Responses from RL model are not surprising to the base model.**
  - **Responses from RL model already exist in the base model's distribution**

$\mathbf{PPL_{Base}}(\boldsymbol{Y}_{\mathbf{Base}}|\boldsymbol{x})$

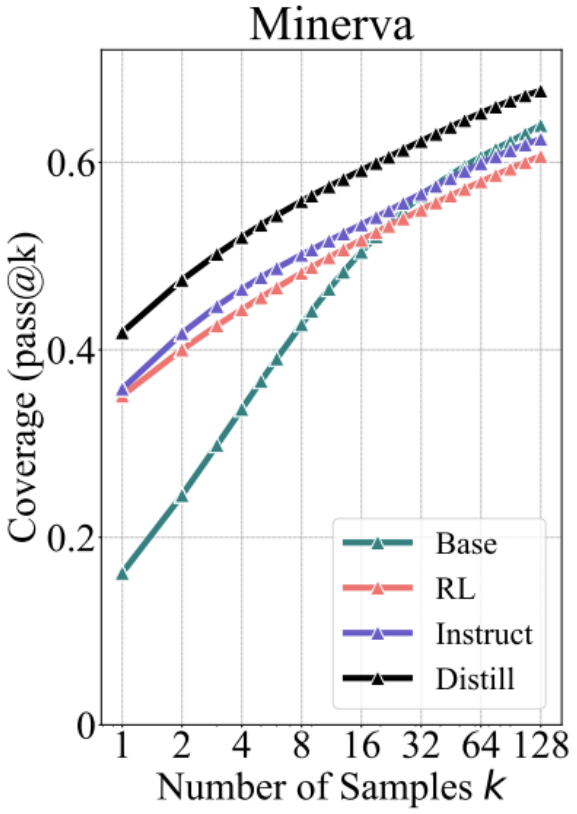
$\mathbf{PPL_{Base}}(\boldsymbol{Y}_{\mathbf{RL}}|\boldsymbol{x})$

PPL lower, likelihood higher



RLVR *v.s.* Distillation

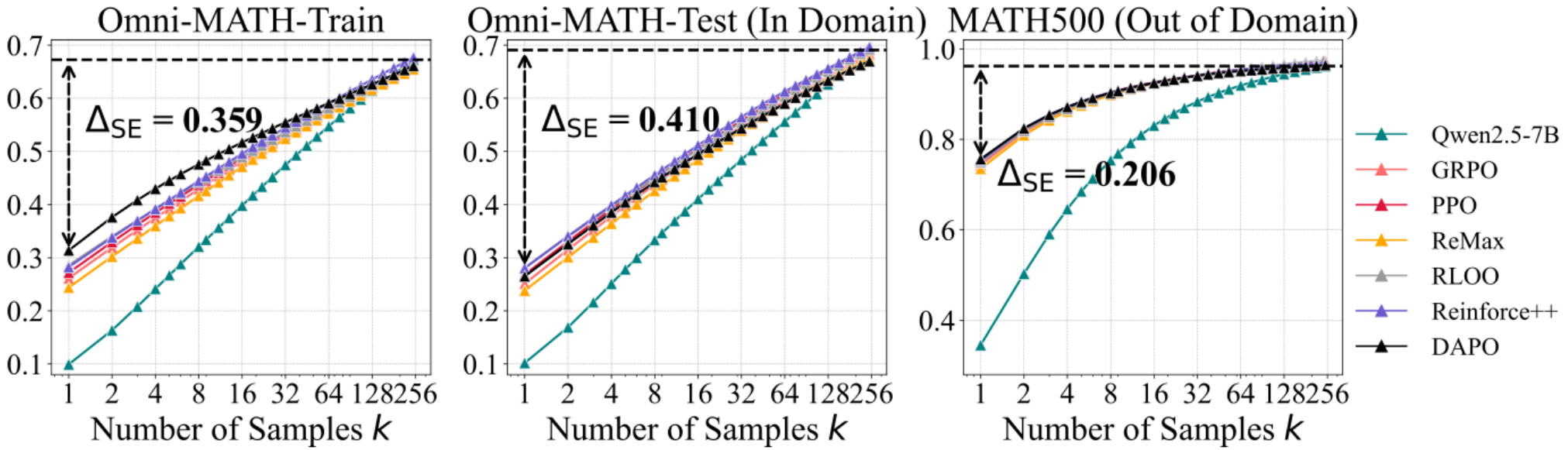
- **Observation: Distilled model consistently outperforms base model as k goes large**
  - **Distillation expands reasoning by injecting new knowledge and reasoning patterns**



Different RLVR Algorithms

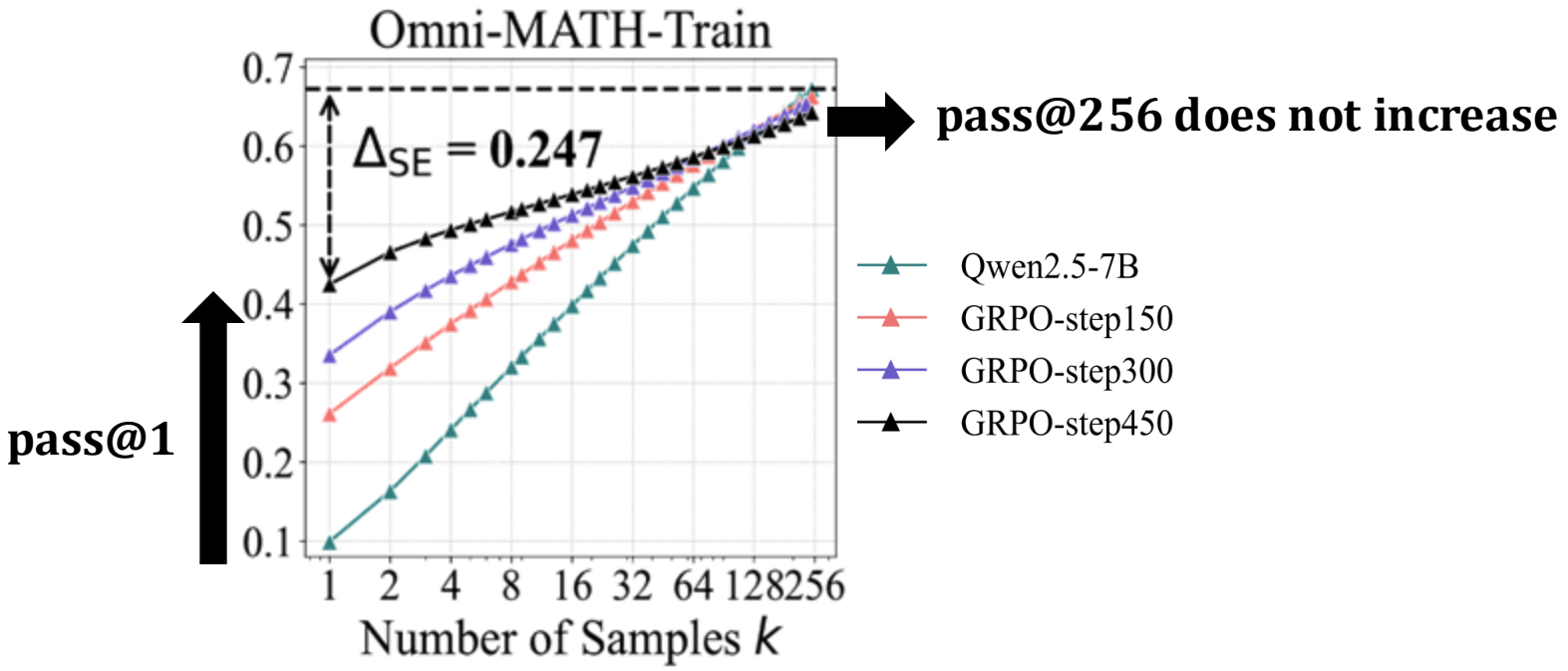
- **Observation 1: Algorithms have slightly different performance**
- **Observation 2: remain far from optimal sampling efficiency**

  Define $\Delta_{SE} = Base\ pass@k\ - RL\ pass@1$ *to measure RL model's sample efficiency*



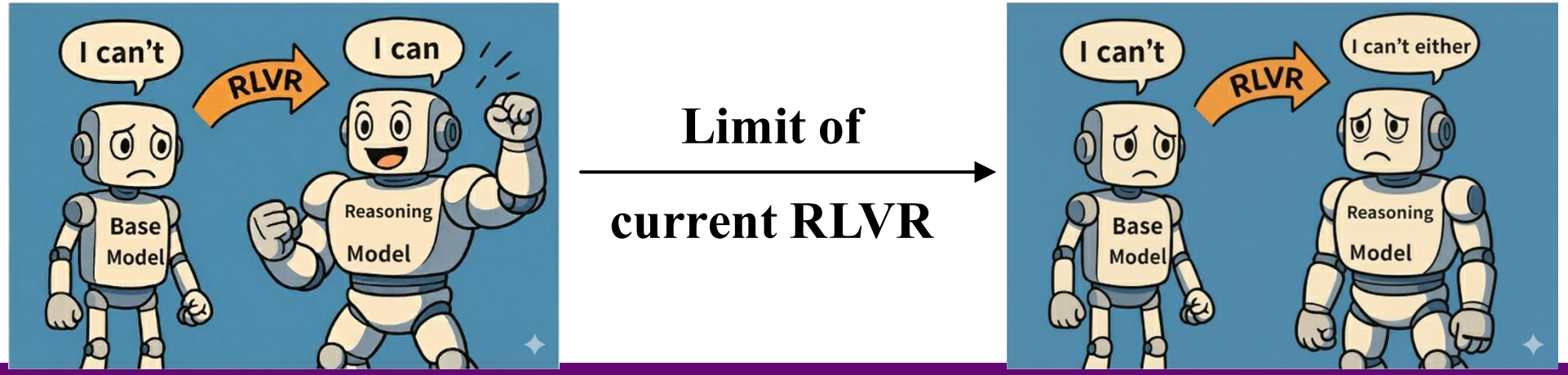
Asymptotic Effects of RL Training

- As RLVR training progresses



Summary

- **Current RLVR in the open-source community may not yet deliver an "*AlphaGo Moment" in LLMs***
  - RLVR is **highly useful in practice:** which mainly comes from the **improvements of sample efficiency**
  - RLVR model is **bounded** by its base model
  - Fully unlock RL's potential to **discover new knowledge / reasoning strategies** remains an open challenge



Discussion: Why RLVR Has Limitations

- **Traditional RL can discover new strategies: AlphaGo's 37th move**

- **Key Differences (RL for Go *vs.* RLVR for LLMs)**
  - **Vast Action Space**: $O(10^{768})$ *vs.* $O(10^{10,000})$
  - **Train from scratch** *vs.* **Pretrained Priors**
    - Pretrain prior guides exploration and make reward possible

Discussion: Why RLVR Has Limitations

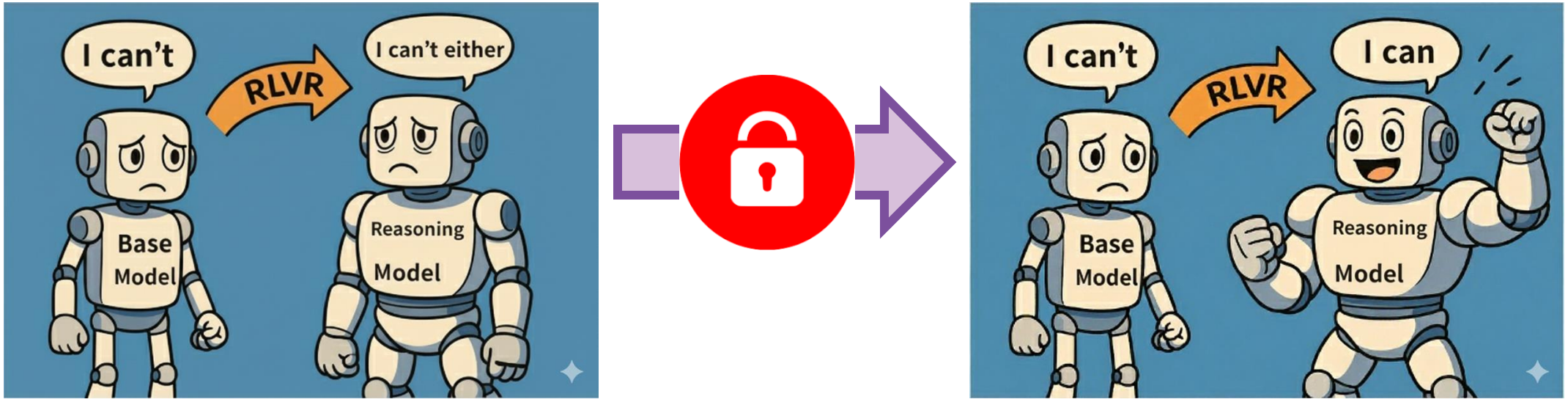
- Traditional RL can discover new strategies: AlphaGo's 37th move

- Key Differences (RL for Go *vs.* RLVR for LLMs)
  - **Vast Action Space**: $O(10^{768})$ *vs.* $O(10^{10,000})$
  - **Train from scratch** *vs.* **Pretrained Priors**
    - Pretrain prior guides exploration and make reward possible

- ***Inability to explore new*** **in this vast action space**
  - Struggle to explore new patterns beyond prior due to:
    - ***Vast action space***
    - ***Naïve exploration (token-level sampling)***
    - ***Sparse reward***

Discussion

How to unlock the potential of RL to discover new



Discussion: How to go beyond the limits of current RLVR

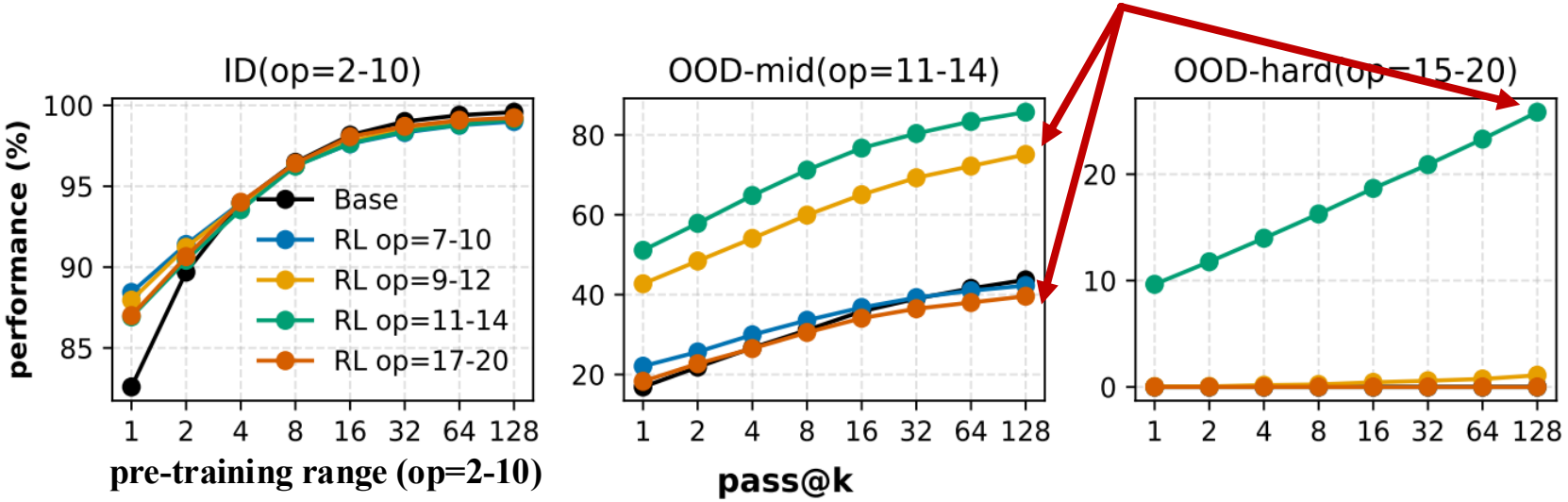
- **Workaround:**
  - **Pretrain/SFT stages inject diversity, and then do RL**

Discussion: How to go beyond the limits of current RLVR

☐ Potential Future Directions:

- ➢ **RL data/env & model scale up**
  - ➢ **Larger models generalize reasoning patterns better**
  - ➢ **Train on diverse data to induce a natural curriculum; progressively learn and compose meta-skills**

RL data should target the model's edge of capability



Zhang, C., Neubig, G., & Yue, X. (2025). On the Interplay of Pre-Training, Mid-Training, and RL on Reasoning Language Models. *arXiv:2512.07783*.

Discussion: How to go beyond the limits of current RLVR

- **Potential Future Directions:**
  - **RL data/env & compute scale up**
  - **Exploration mechanism**
    - **Beyond inefficient sampling; evolve trajectories from past explorations (*e.g.*, AlphaEvolve).**

Discussion: How to go beyond the limits of current RLVR

- **Potential Future Directions:**
  - RL data/env & compute scale up
  - Exploration mechanism (*i.e.*, AlphaEvolve)
  - **Process reward & value network**
    - **Give intermediate feedback before we get a complete solution**

Discussion: How to go beyond the limits of current RLVR

- **Potential Future Directions:**
  - RL data/env & compute scale up
  - Exploration mechanism (*i.e.*, AlphaEvolve)
  - Process reward & value network
  - **Agent interaction with tools and external info**
    - **rich input is needed for creation**

Discussion: How to go beyond the limits of current RLVR

- **Potential Future Directions:**
  - **RL data/env & compute scale up**
  - **Exploration mechanism (*i.e.*, AlphaEvolve)**
  - **Process reward & value network**
  - **Agent interaction with tools and external info**

} **towards unlocking full potential of RL**

Thanks to all the collaborators

Yang Yue (乐洋)
Seeking an internship
in North America

Zhiqi Chen
Undergrad
Seeking summer research

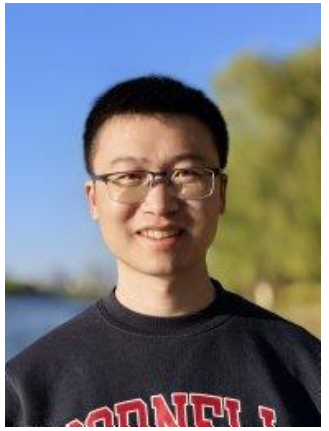
Rui Lu
Ph.D.
On the job market

Andrew Zhao

Zhaokai Wang

Yang Yue (乐阳)

Shiji Song



Gao Huang

清華大學
Tsinghua University

Thanks for your listening!

See our web for more discussions and Q&A

https://limit-of-rlvr.github.io