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

Yang Yue[1*†], Zhiqi Chen[1*], Rui Lu[1], Andrew Zhao[1], Zhaokai Wang[2], Yang Yue[1], Shiji Song[1], and Gao Huang[1‡]

[1] Tsinghua University [2] Shanghai Jiao Tong University

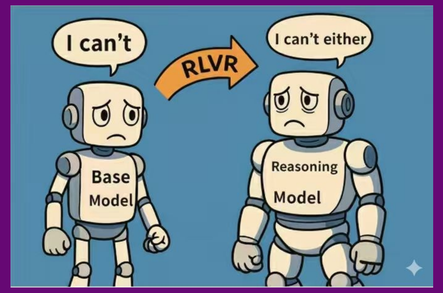
[*] Equal Contribution [†] Project Lead [‡] Corresponding Author

paper homepage

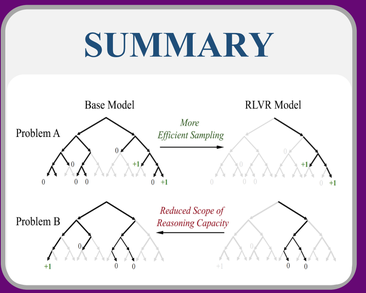
NEURAL INFORMATION PROCESSING SYSTEMS

oral

RLVR rarely expands reasoning boundary, mainly improves sampling efficiency!

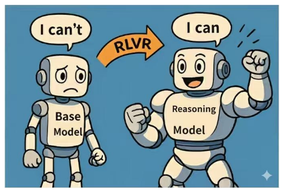


SUMMARY

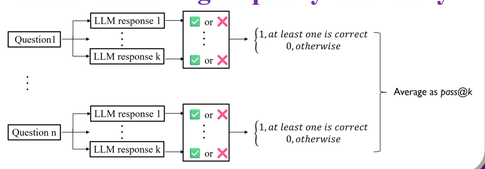


1 Background & Preliminaries

◆**Common Belief for RLVR**



◆**[pass@k] A metrics for LLM Reasoning Capacity Boundary**



2 Experiment Setup

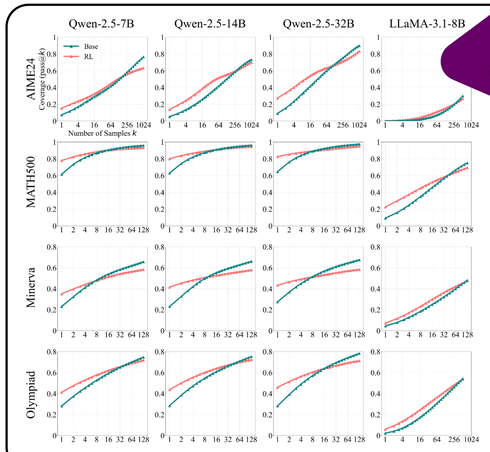
Task	Start Model	RL Framework	RL Algorithm(s)	Benchmark(s)
Mathematics	LLaMA-3.1-8B Qwen-2.5-7B/14B/32B-Base Qwen-2.5-Math-7B	SimpleRLZoo Oat-Zero	GRPO	GSM8K, MATH500 Minerva, Olympiad AIME24, AMC23
Code Generation	Qwen-2.5-7B-Instruct	Code-R1	GRPO	LiveCodeBench HumanEval+
Visual Reasoning	Qwen-2.5-VL-7B-Instruct	VeRL	PPO	MathVista MathVision
Deep Analysis	Qwen-2.5-7B-Base Qwen-2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-7B	VeRL	PPO, GRPO Reinforce++ RLOO, ReMax, DAPO	Omni-Math-Rule MATH500

Base vs RLVR

Keep prompt template for base model identical to RLVR template

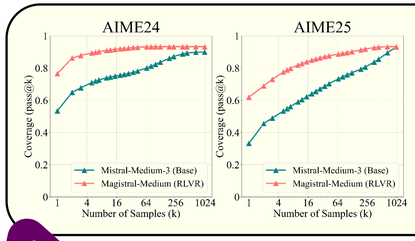
3 Key Findings

◆**RL-trained models perform worse than base models in pass@k at large k values**



RLVR for Mathematical Reasoning

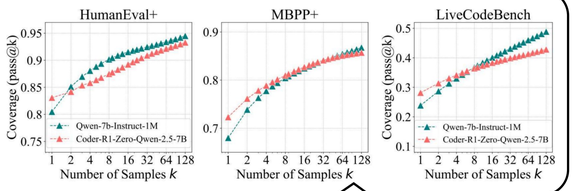
1. **Increased** Likelihood of Correct Samples at **small k**
2. **Decreased** Coverage of Solvable Problems at **large k**



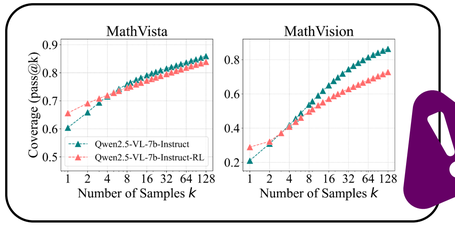
Similar Findings!!!

RLVR for Mathematical Reasoning in Near-frontier Models

RLVR for Code Generation



Unit tests are hard to pass by **GUESSWORK!**

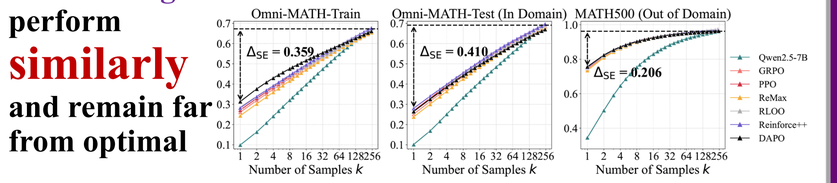


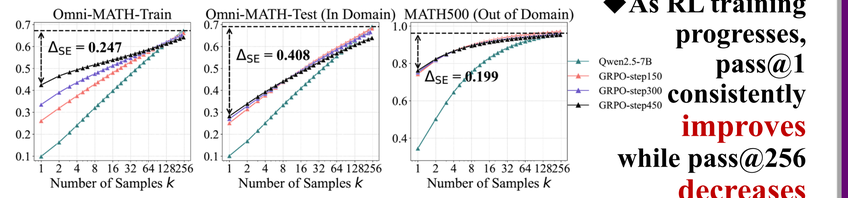
RLVR for Visual Reasoning

4 Deep Analysis

◆**RLVR algorithms perform similarly and remain far from optimal**

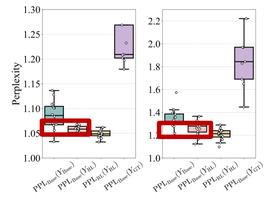
Define $\Delta_{SE} = Base\ pass@k - RL\ pass@1$.





◆**As RL training progresses, pass@1 consistently improves while pass@256 decreases**

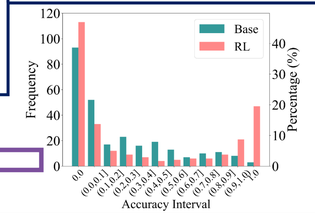
◆**Reasoning Patterns Maybe Already Present in Base Models**



Responses from ***RL-trained*** models are ***highly likely*** to be generated by the ***base*** model.

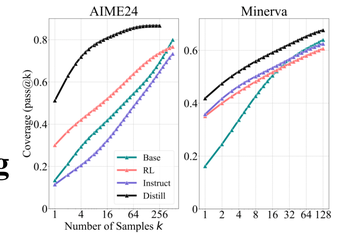
Base	SimpleRLZoo	AIME24	MATH500
✓	✓	63.3%	92.4%
✓	✗	13.3%	3.6%
✗	✓	0.0%	1.0%
✗	✗	23.3%	3.0%

The RLVR Model's Solvable-Problem Coverage is an ***Approximate Subset*** of the Base Model's Coverage.



The improvement in average scores, driven by improving ***sampling efficiency*** on ***already solvable*** problems.

◆**RLVR and distillation are fundamentally different: distillation expands reasoning by injecting new knowledge.**



Potential Future Directions

1. RL data & compute scale up!
2. Agent with tools & external info
3. Exploration mechanism (AlphaEvolve)
4. Process reward & better credit assignment

AbsoluteZero: A prior work on scaling learning tasks

80/20: The inspiration of token-level entropy