ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models

刁诗哲
英伟达 研究员

Team Members

ProRL: Prolonged Reinforcement Learning Expands Reasoning Boundaries in Large Language Models

Mingjie Liu **Shizhe Diao** **Ximing Lu** **Jian Hu** **Xin Dong**
Yejin Choi **Jan Kautz** **Yi Dong**
NVIDIA
{mingjiel, sdiao, ximingl, jianh, xind, yejinc, jkautz, yidong}@nvidia.com

2025-7-2

Prolonged Reinforcement Learning for Diverse LLM Reasoning

A. Contributors

Mingjie Liu, Shizhe Diao, Jian Hu, Ximing Lu, Xin Dong, Hao Zhang, Alexander Bukharin, Shaokun Zhang, Jiaqi Zeng, Makesh Narsimhan Sreedhar, Gerald Shen, David Mosallanezhad, Di Zhang, Jonas Yang, June Yang, Oleksii Kuchaiev, Guilin Liu, Zhiding Yu, Pavlo Molchanov, Yejin Choi, Jan Kautz, Yi Dong

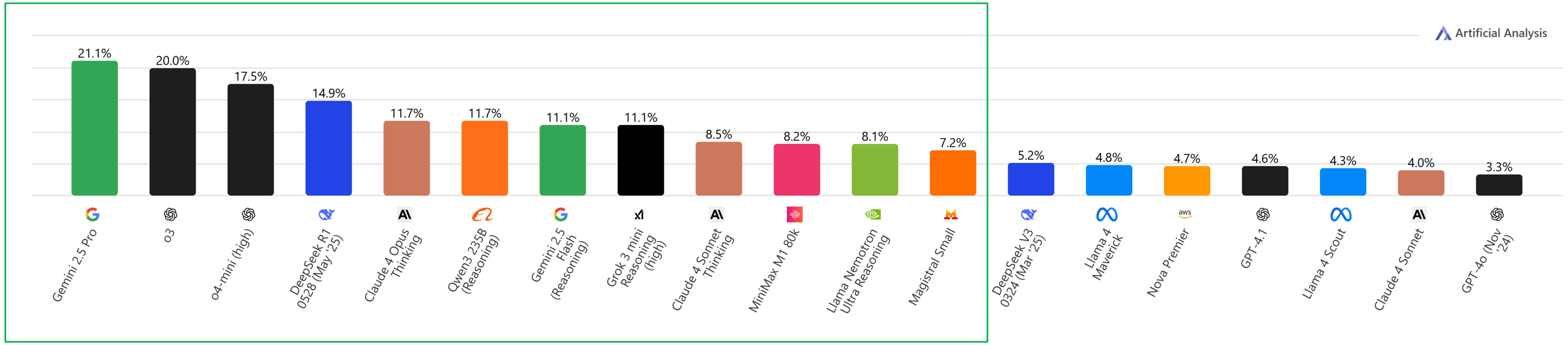
2025 A Reasoning Model Year

A Paradigm Shift

- Explosion of reasoning models
  - Commercial models
    - OpenAI's o-series, Claude 3.7/4.0 Sonnet, and Gemini 2.5 Pro
  - Open-Source models
    - DeepSeek-R1, Qwen QwQ, Qwen 3 family, NVIDIA Llama Nemotron Family
- What Sets Reasoning Models Apart?
  - Extended thinking time using long chain-of-thought and stepwise processing
  - Trained via reinforcement learning techniques
  - State-of-the-art performance on challenging reasoning benchmarks

Humanity's Last Exam Benchmark Leaderboard: Results

Independently conducted by Artificial Analysis



The Ultimate Question For Developing Reasoning Model

- Does reinforcement learning **truly unlock new reasoning capabilities** from a base model, or does it merely **optimize the sampling efficiency** of solutions already embedded in the base model (temperature distillation)?

Claims: No Acquired Capabilities Beyond Base Models

Decomposing Elements of Problem Solving: What "Math" Does RL Teach?

Tian Qin[1]* Core Francisco Park[1]*
Mujin Kwun[1,2] Aaron Walsman[1,2] Eran Malach[1,2] Nikhil Anand[1,2]
Hidenori Tanaka[1]† David Alvarez-Melis[1,2]†
[1]Harvard University [2]Kempner Institute

Assessing Diversity Collapse in Reasoning

Xingyu Dang, Christina Baek, J Zico Kolter, Aditi Raghunathan

Published: 07 Mar 2025, Last Modified: 07 Mar 2025 SSI-FM Poster Everyone Revisions BibTeX CC BY 4.0
Keywords: LLM, reasoning, supervised finetuning, reinforcement learning, decoding strategy

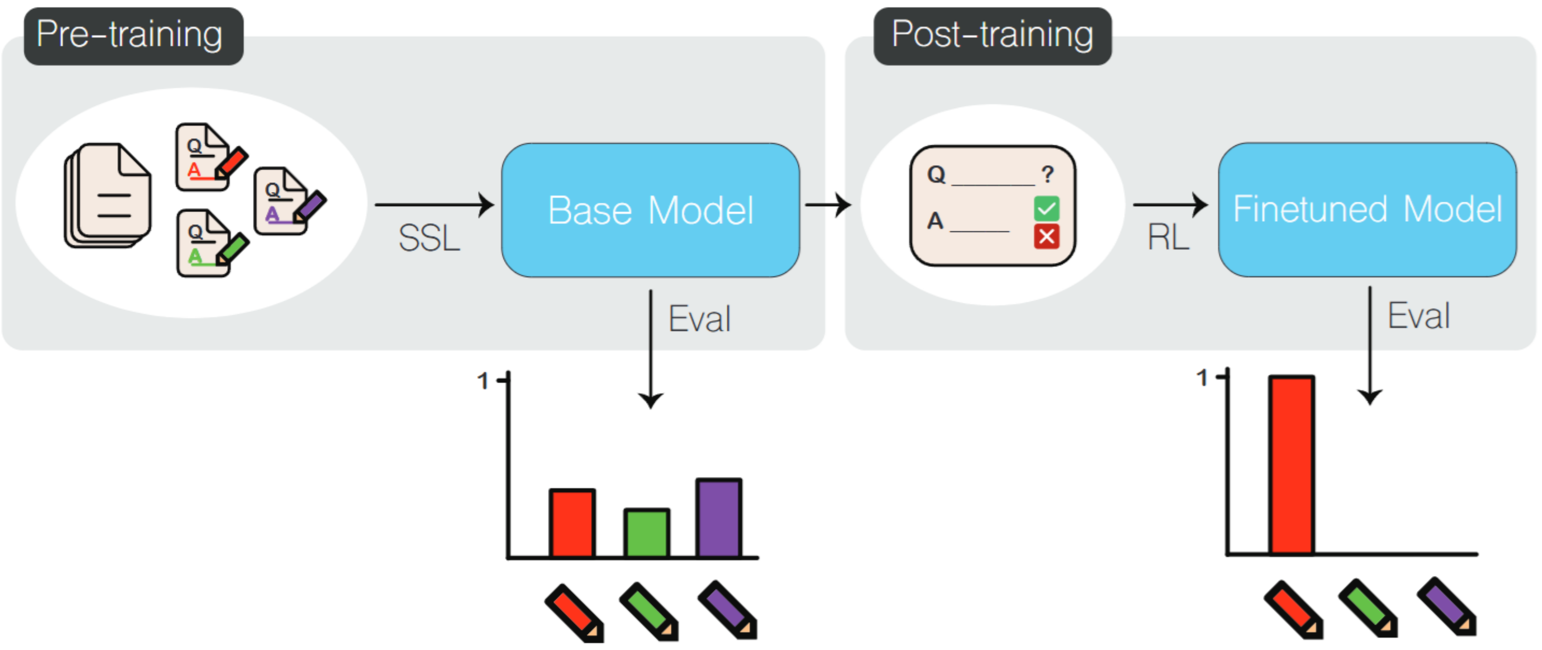
Echo Chamber: RL Post-training Amplifies Behaviors Learned in Pretraining

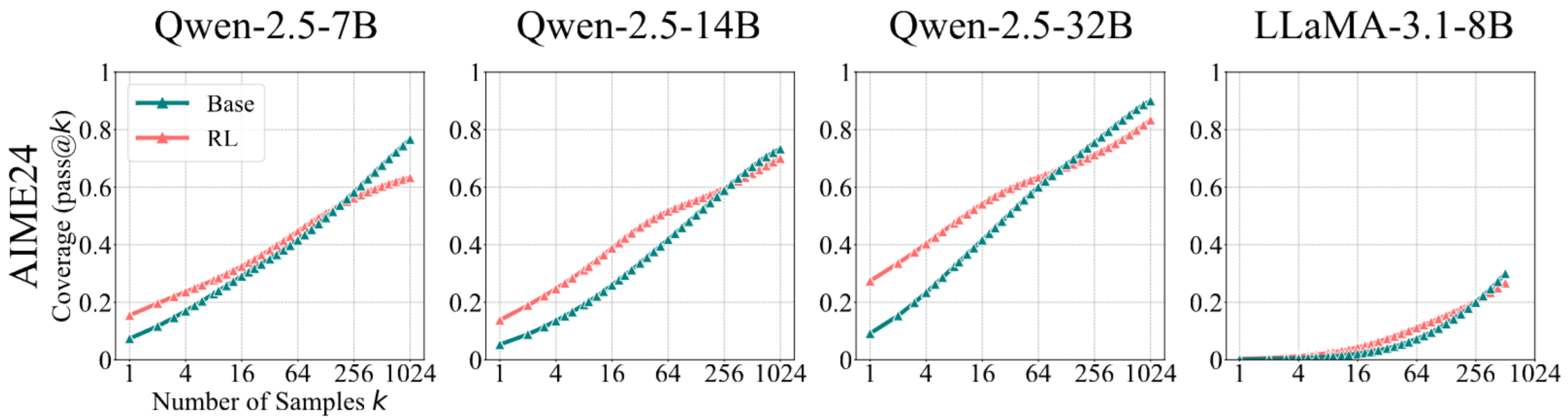
Rosie Zhao* Harvard University Kempner Institute
Alexandru Meterez* Harvard University Kempner Institute
Sham Kakade Harvard University Kempner Institute
Cengiz Pehlevan Harvard University Kempner Institute
Samy Jelassi† Harvard University
Eran Malach† Harvard University Kempner Institute

LEAP *May 19, 2025*

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]LeapLab, Tsinghua University [2]Shanghai Jiao Tong University
* Equal Contribution † Project Lead ✉ Corresponding Author





Is LLM RL doomed?

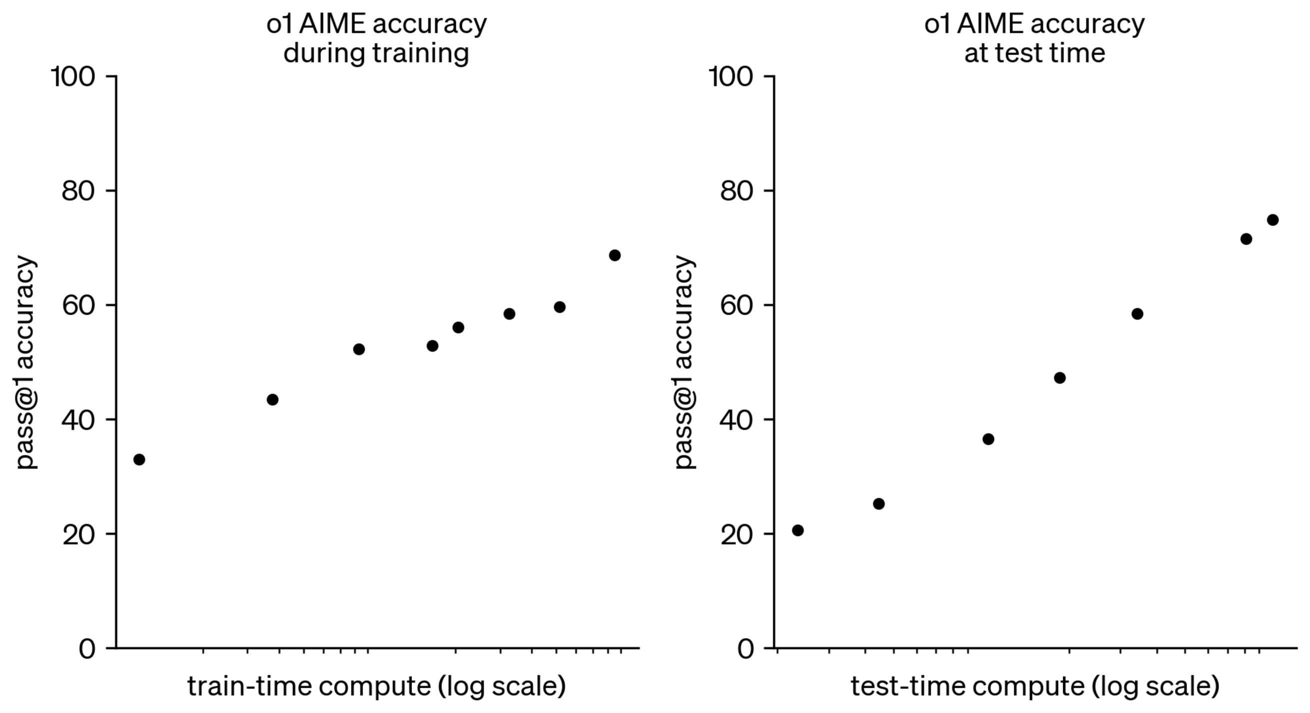
- Temperature distillation is boring, what about superhuman intelligence?

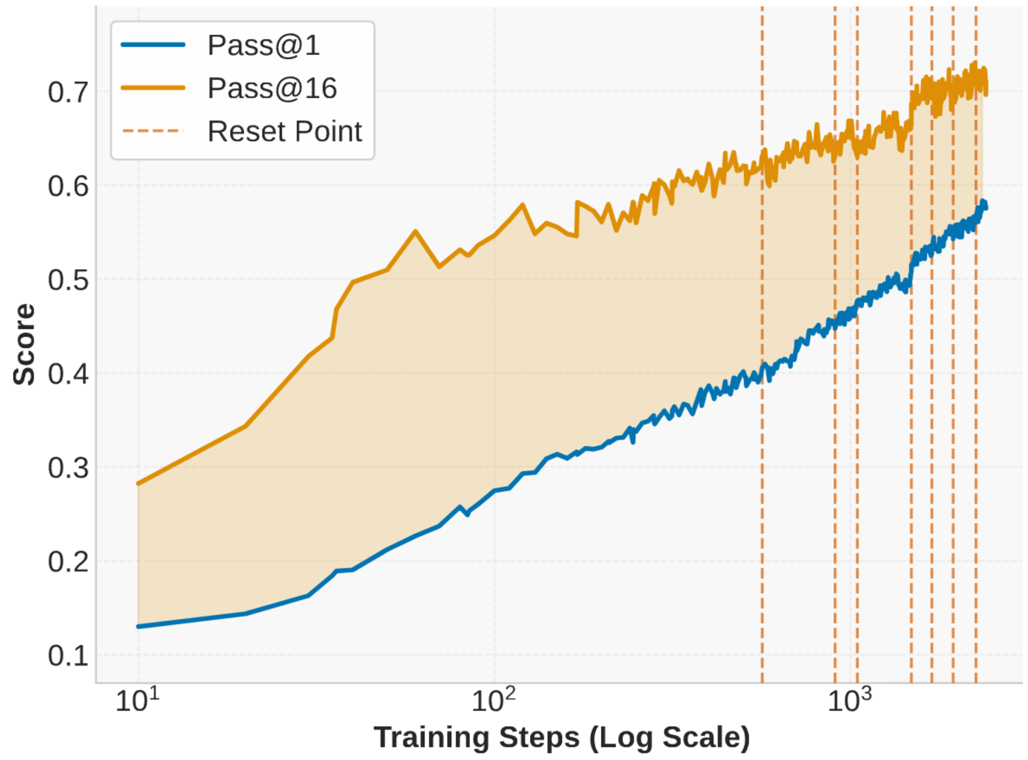
- What's common in previous studies:
  - An overreliance on specialized domains like mathematics that is overtrained during both pre-training and post-training phases
    - Can someone be creative if he only trains on the tasks that he is already good at?
  - The premature termination of RL training, typically no more than hundreds of step
    - Can someone discover new ideas if he is only allowed to explore a new area for a short amount of time?
- Luckily, our study find evidences that models learns new capabilities beyond the base model.
  - Implication: potentially achieve superhuman intelligence just by running RL.

Our Philosophy to Train Reasoning Models

In contrast to previous studies, ProRL is different

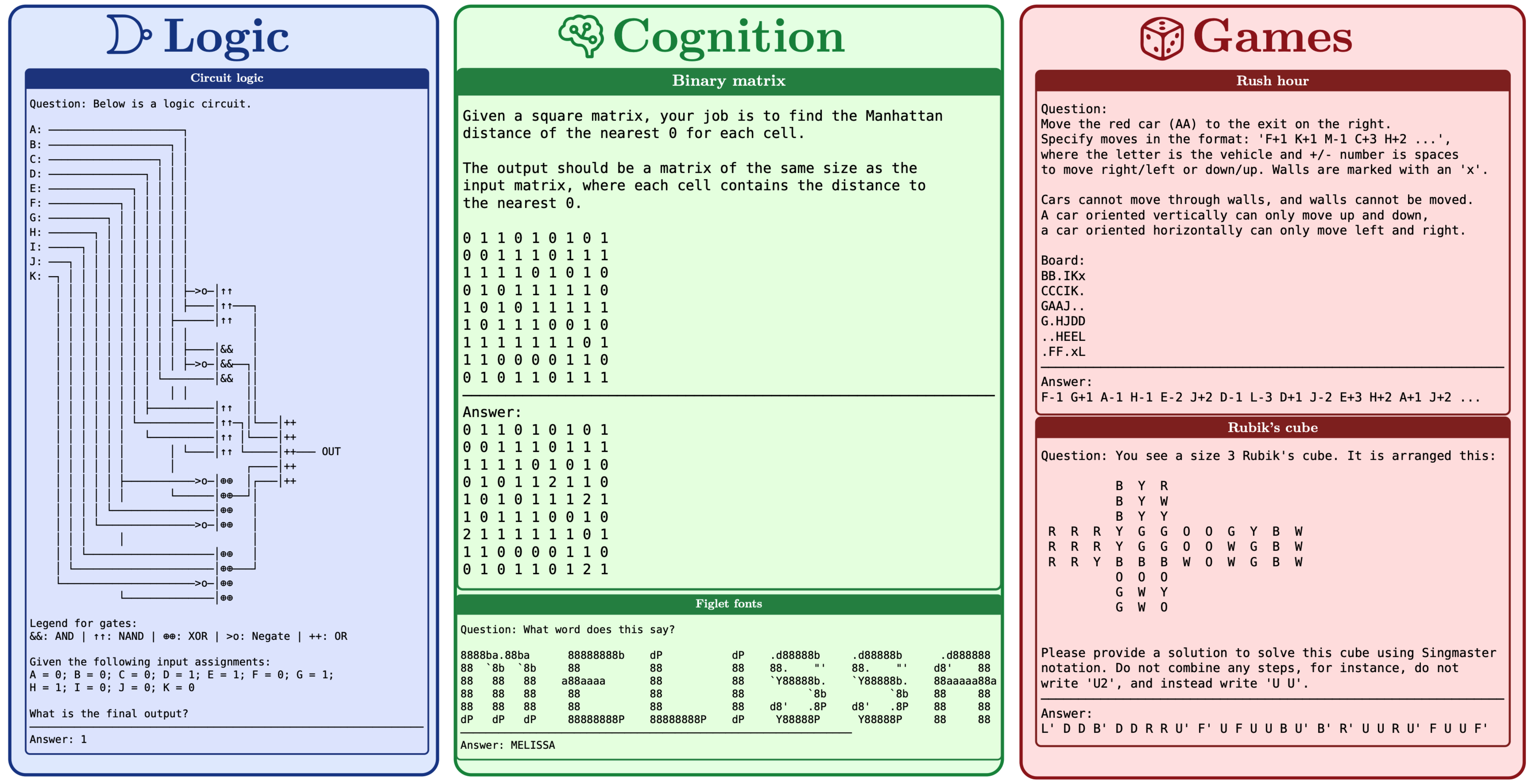
- Prolonged Reinforcement Learning, scale the RL training





- Using diversified novel reasoning tasks

Data Type	Reward Type	Quantity	Data Source
Math	Binary	40k	DeepScaleR Dataset
Code	Continuous	24k	Eurus-2-RL Dataset
STEM	Binary	25k	SCP-116K Dataset
Logical Puzzles	Continuous	37k	Reasoning Gym
Instruction Following	Continuous	10k	Llama-Nemotron



Outlines

- How to achieve Prolonged reinforcement learning
- Reasoning model training Results - ProRL produced SOTA 1.5B reasoning model
- Analysis on the results to address the question whether RL expand the reasoning boundary

How to Achieve ProRL

- We use GRPO RL algorithm with DAPO tricks
  - Dynamic sampling
  - Decoupled clip high and low

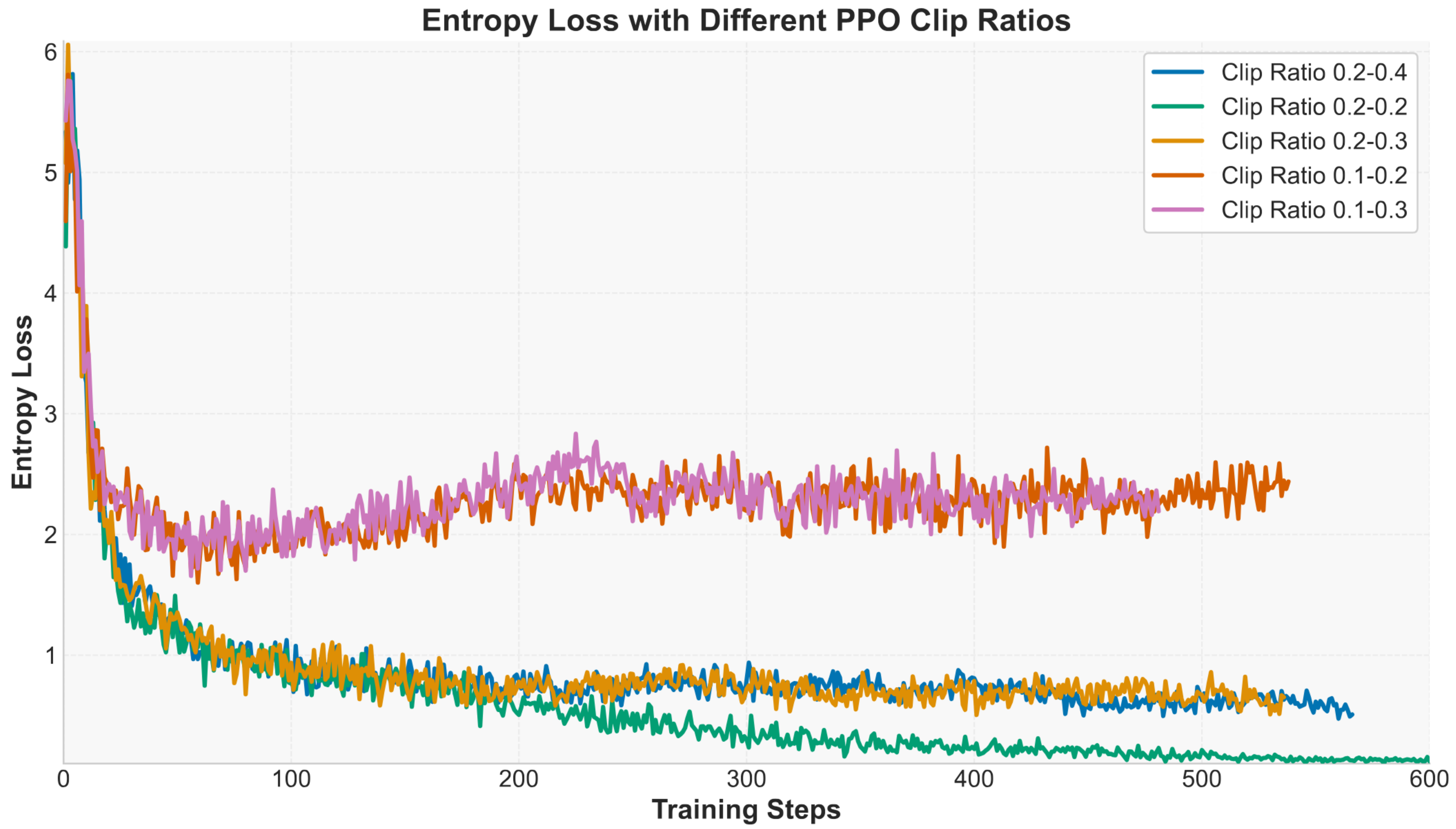
$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\min\left(r_\theta(\tau)A(\tau), \quad \text{clip}(r_\theta(\tau), 1-\epsilon, 1+\epsilon)A(\tau)\right)\right]$$

$$\text{clip}(r_\theta(\tau), 1-\epsilon_{low}, 1+\epsilon_{high})$$

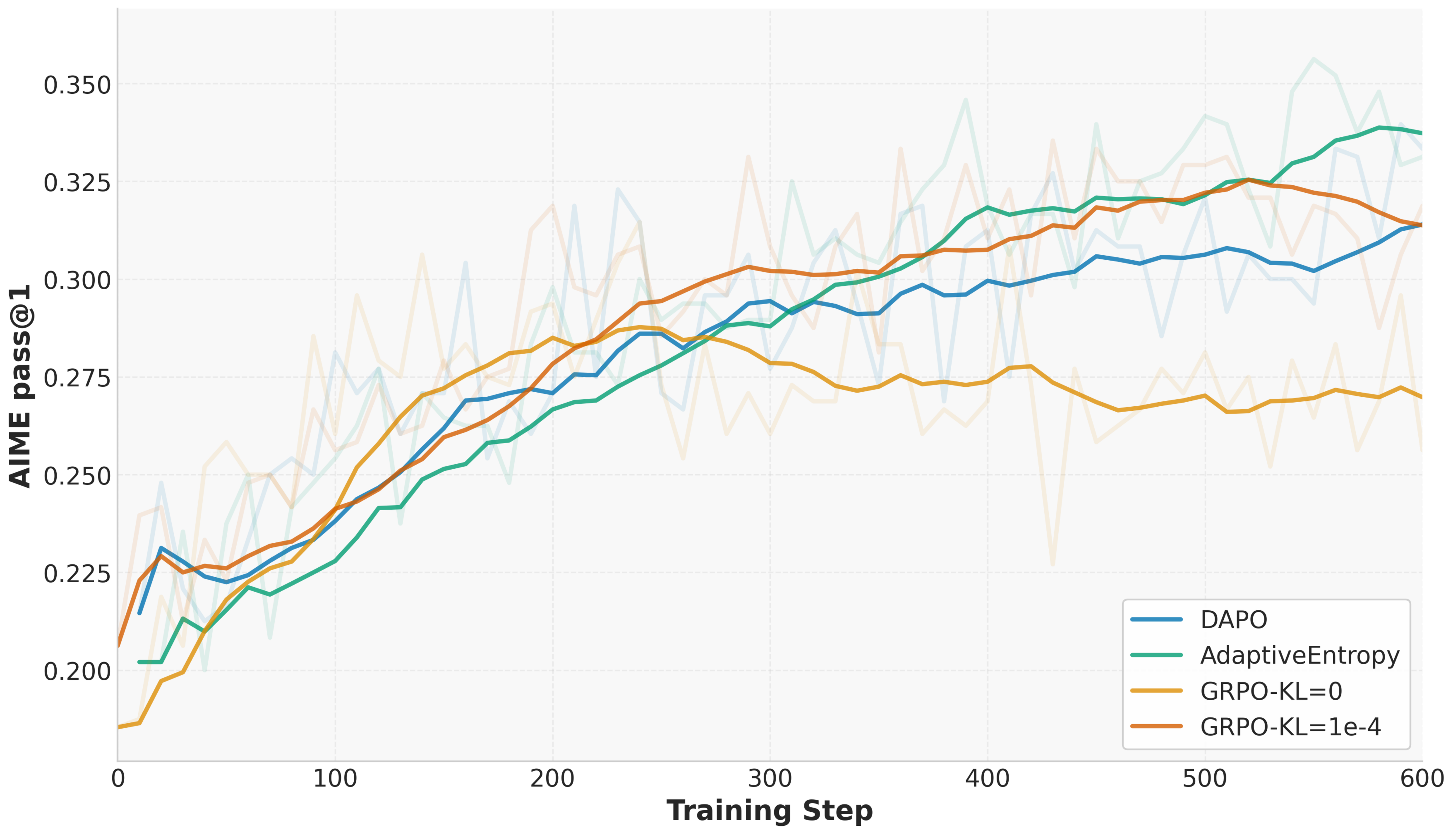
- Balance the exploration and exploitation
  - Sustainable entropy
- Resetting the reference policy and optimizer states
  - No bounded exploration

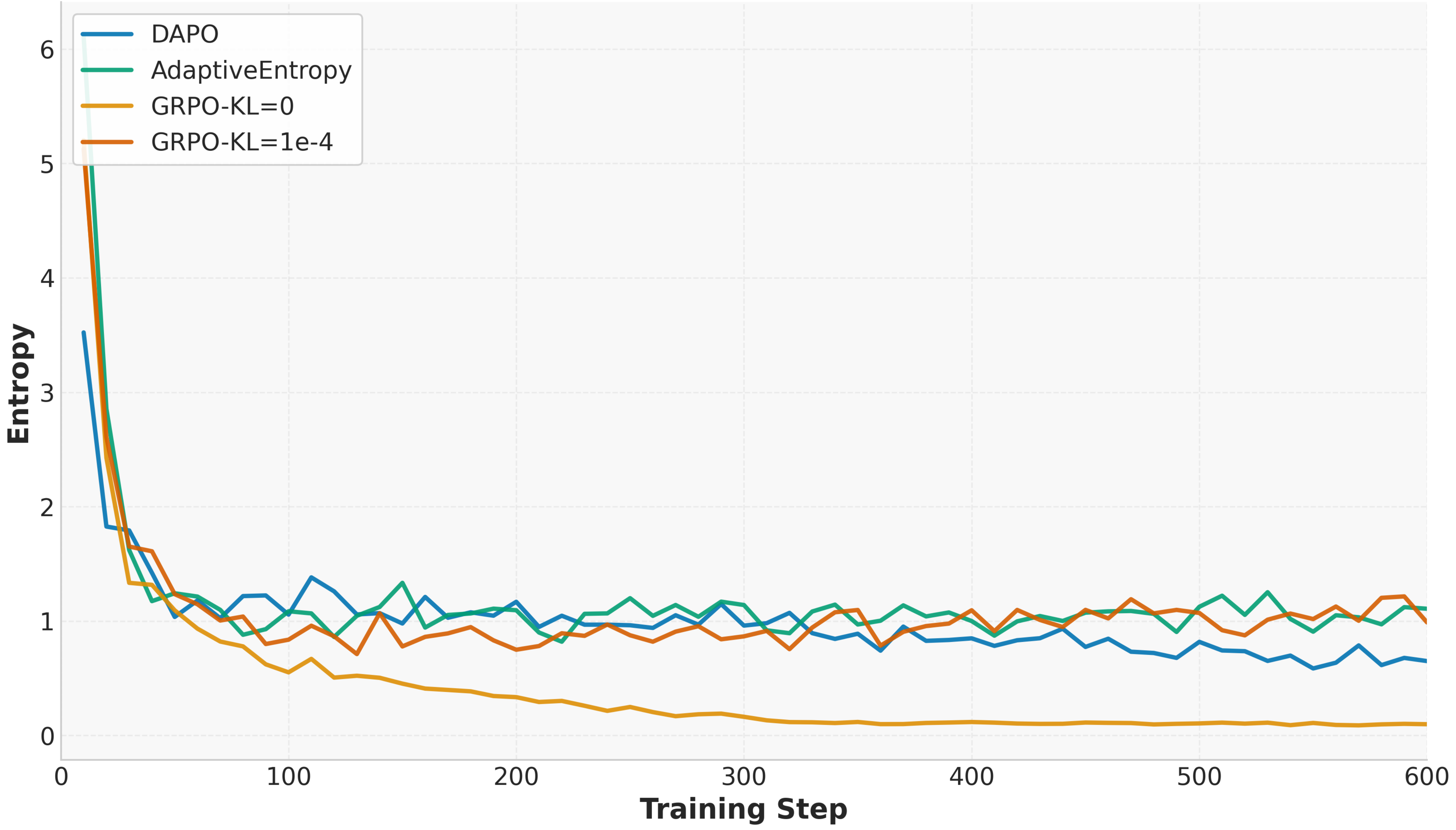
Balance the Exploration and Exploitation

- Entropy trending up or down is bad
  - It is not sustainable for prolonged training, the same reasoning that exploding/vanishing gradient is bad in training.



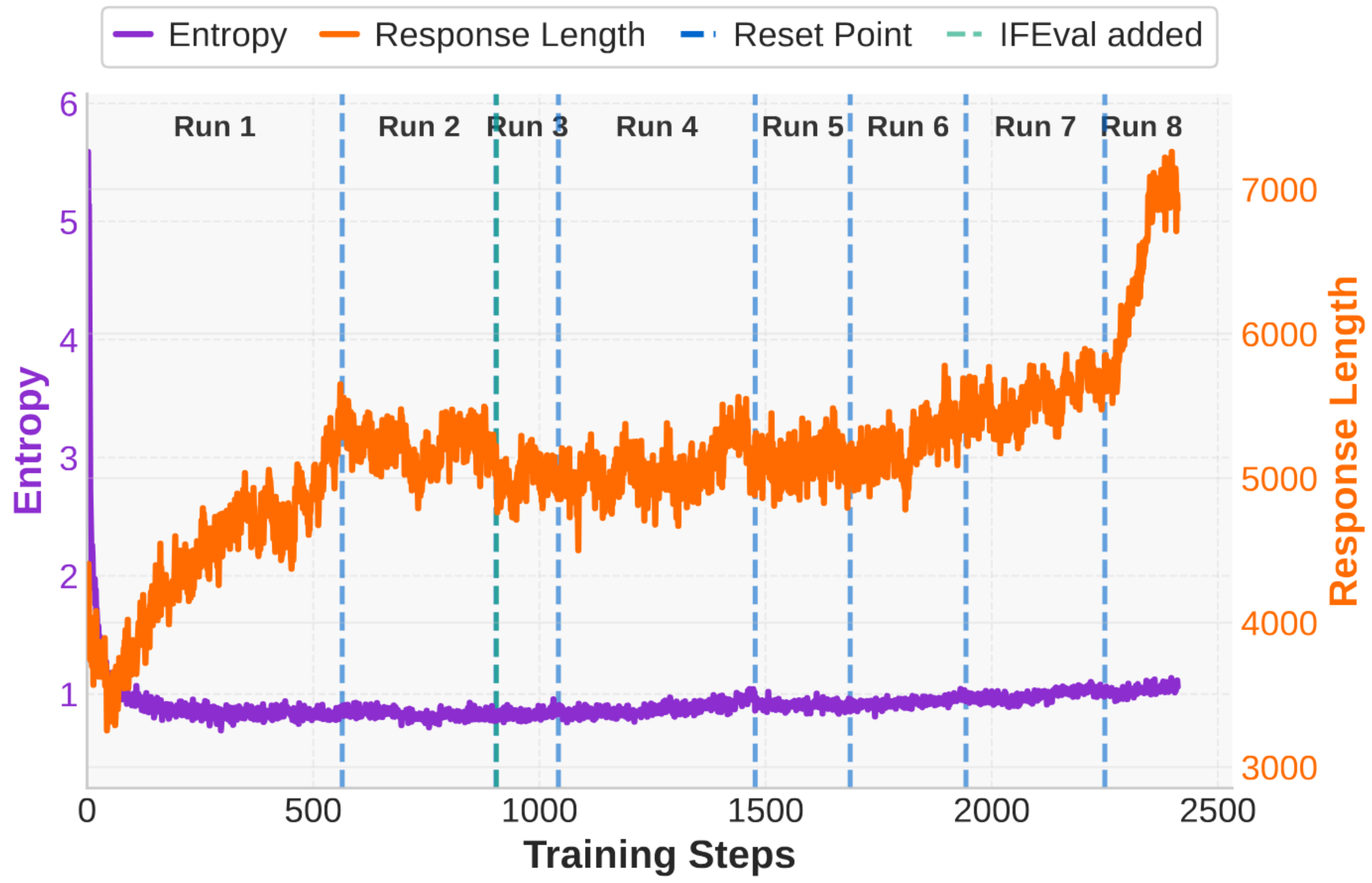
We need KL regularization to maintain constant entropy





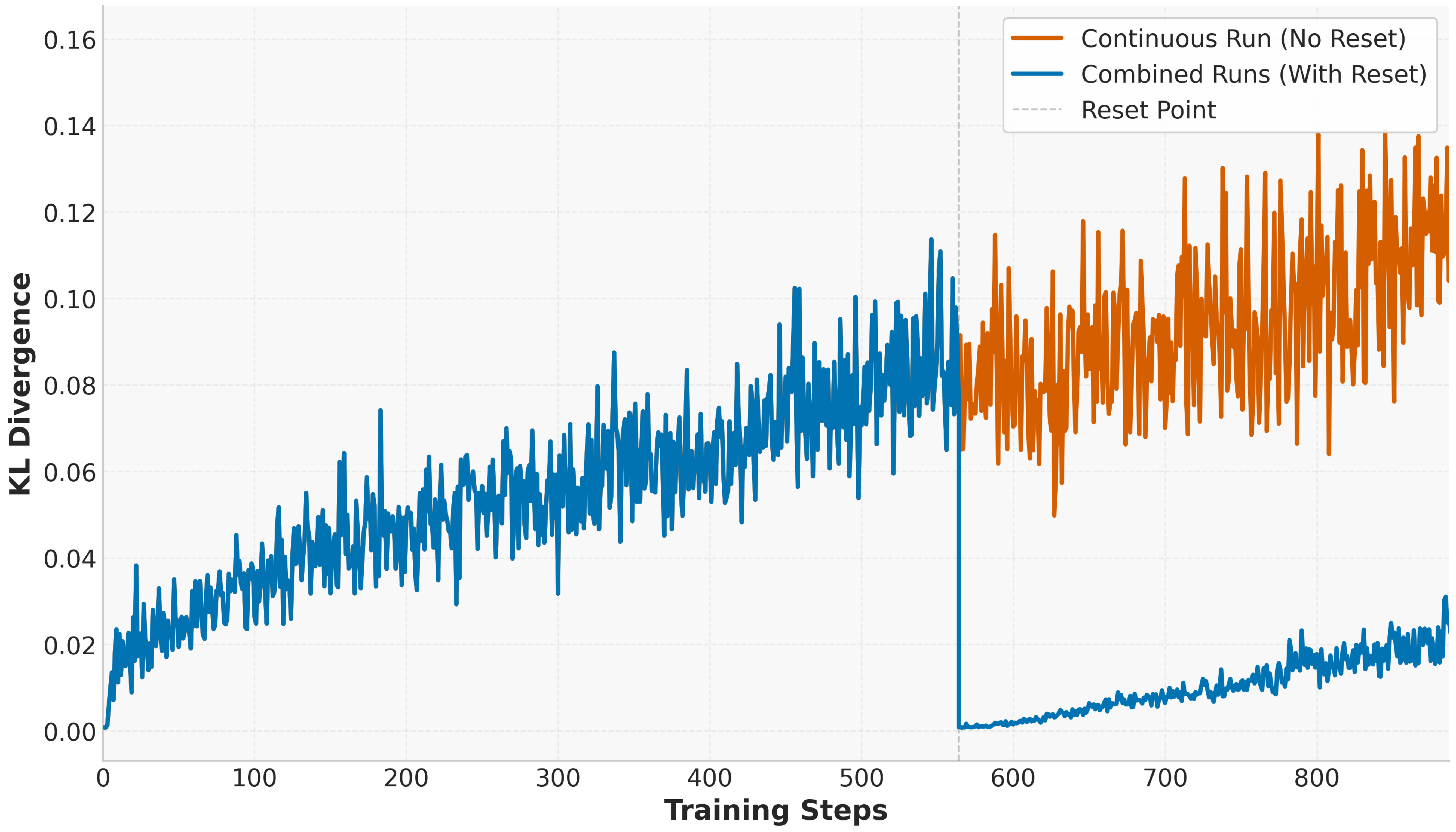
Stable Entropy

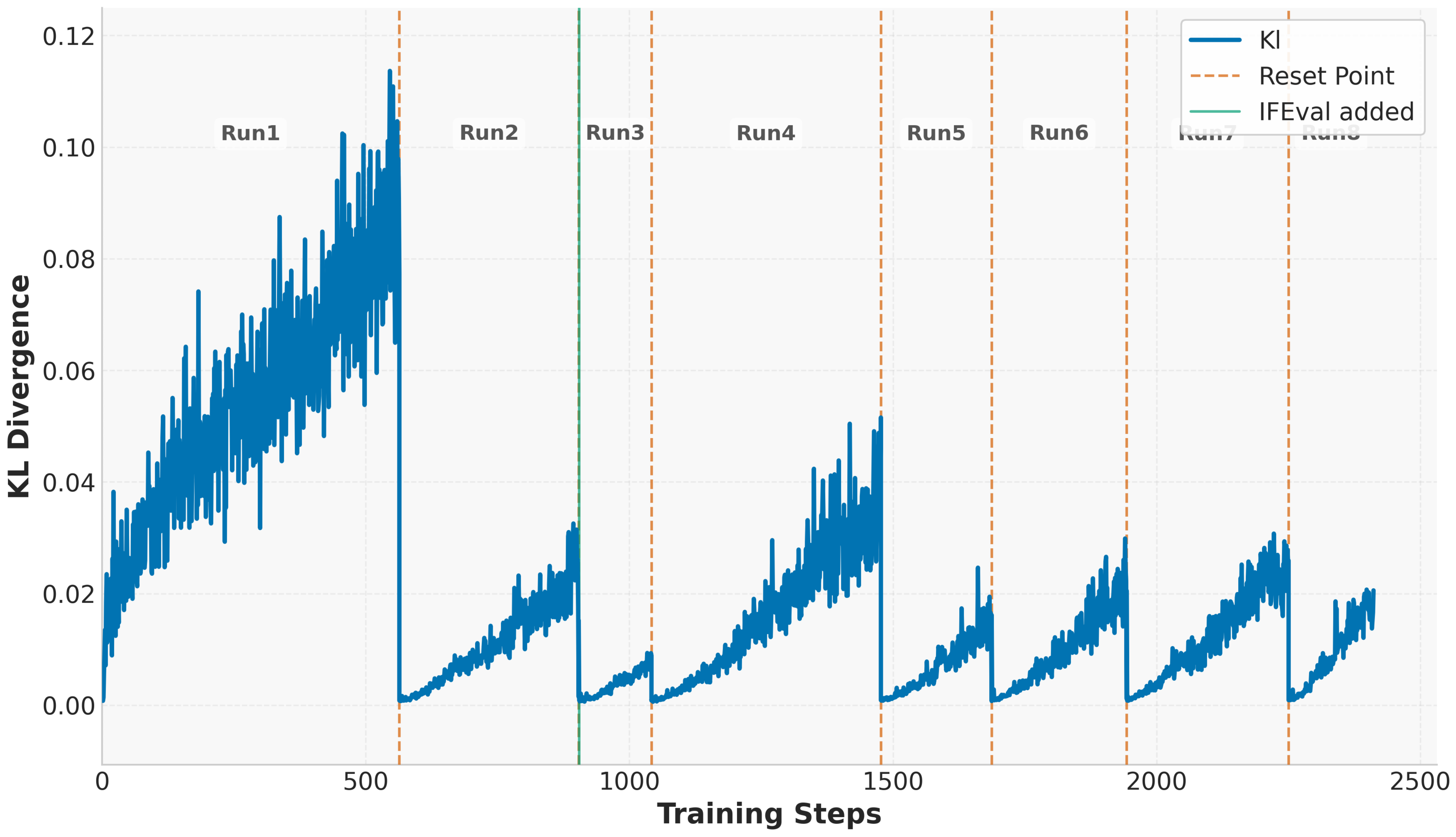
A good balance between exploration and exploitation



Importance of Reference Model Resetting

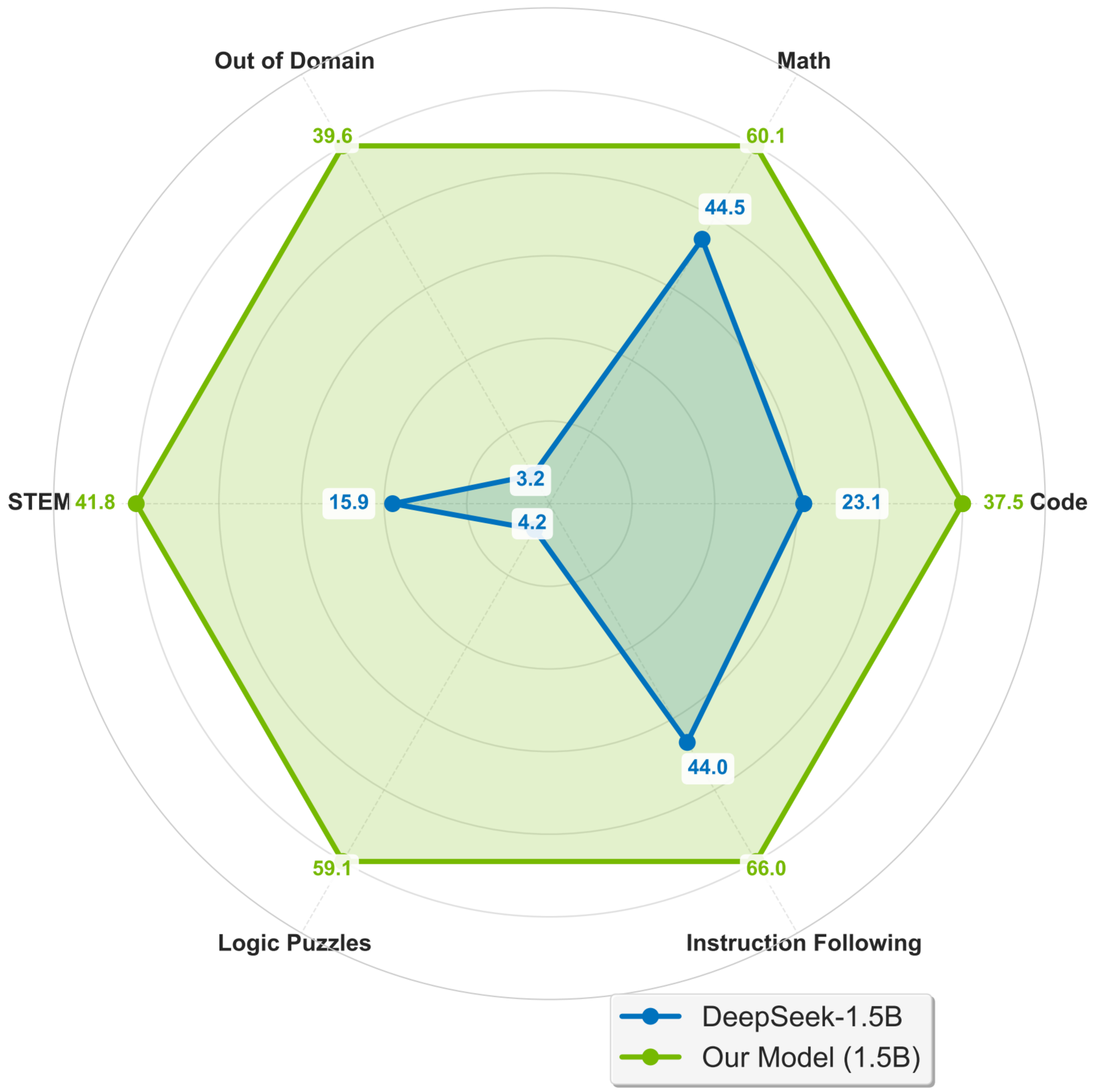
Stabilize the Training

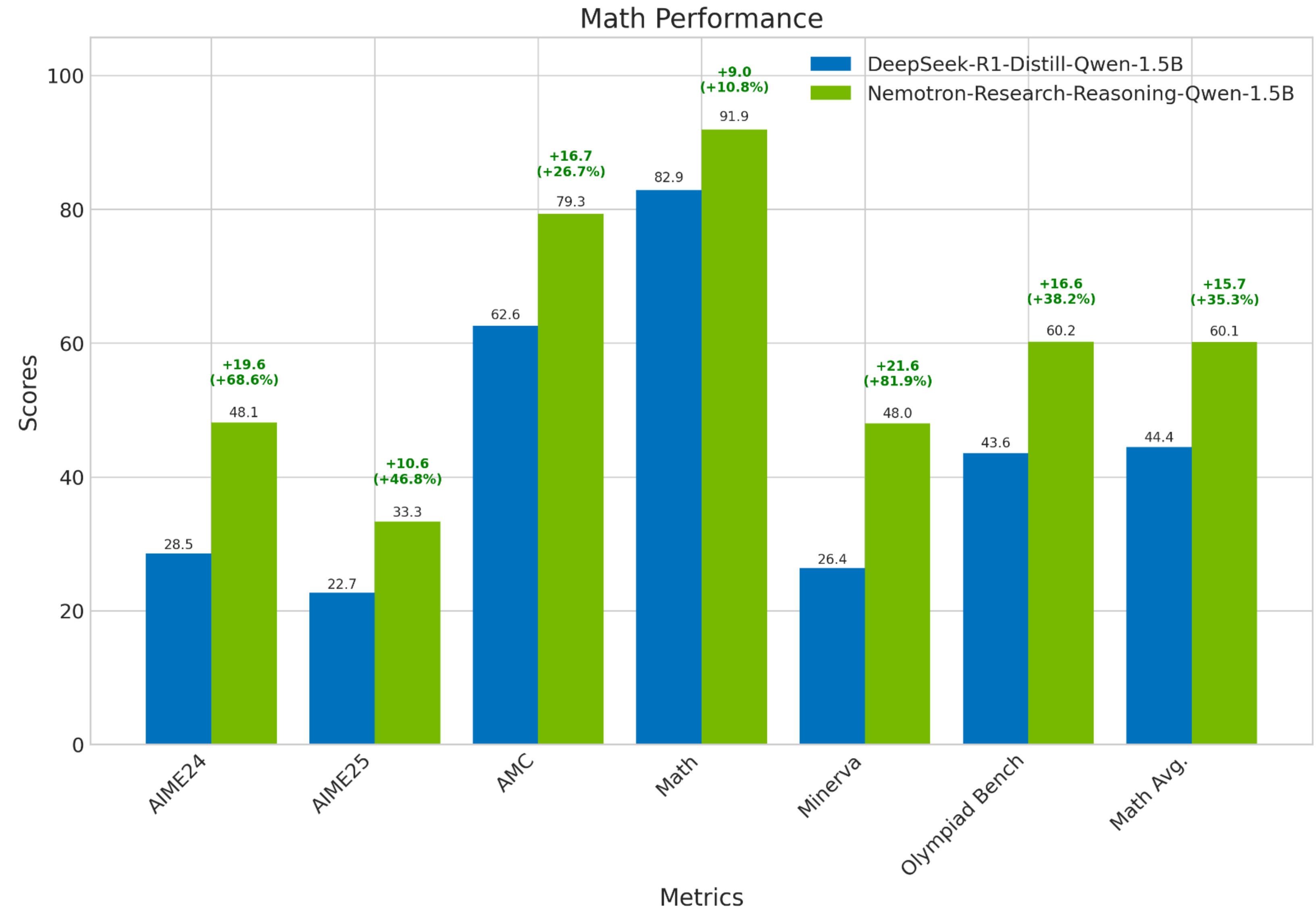


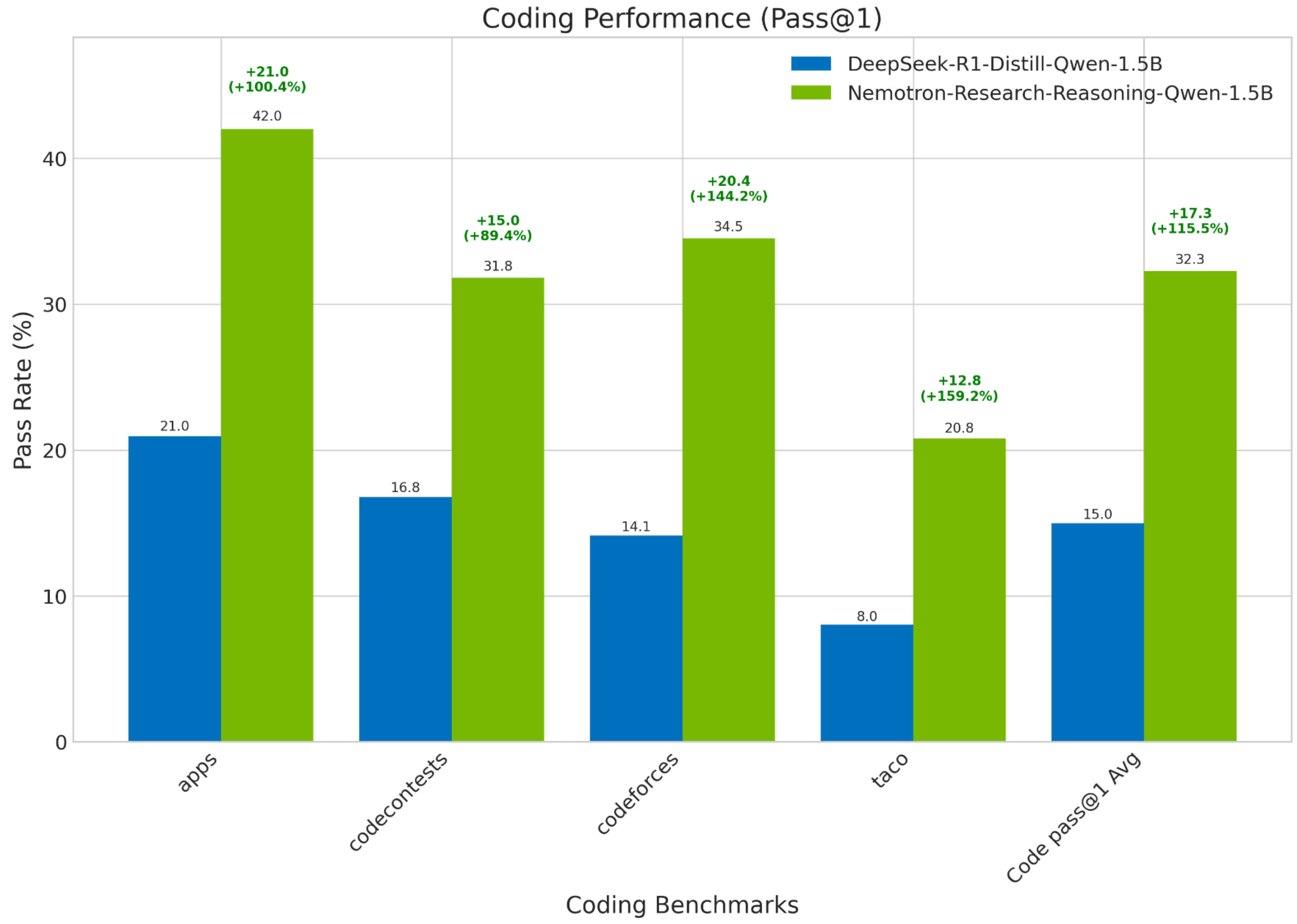


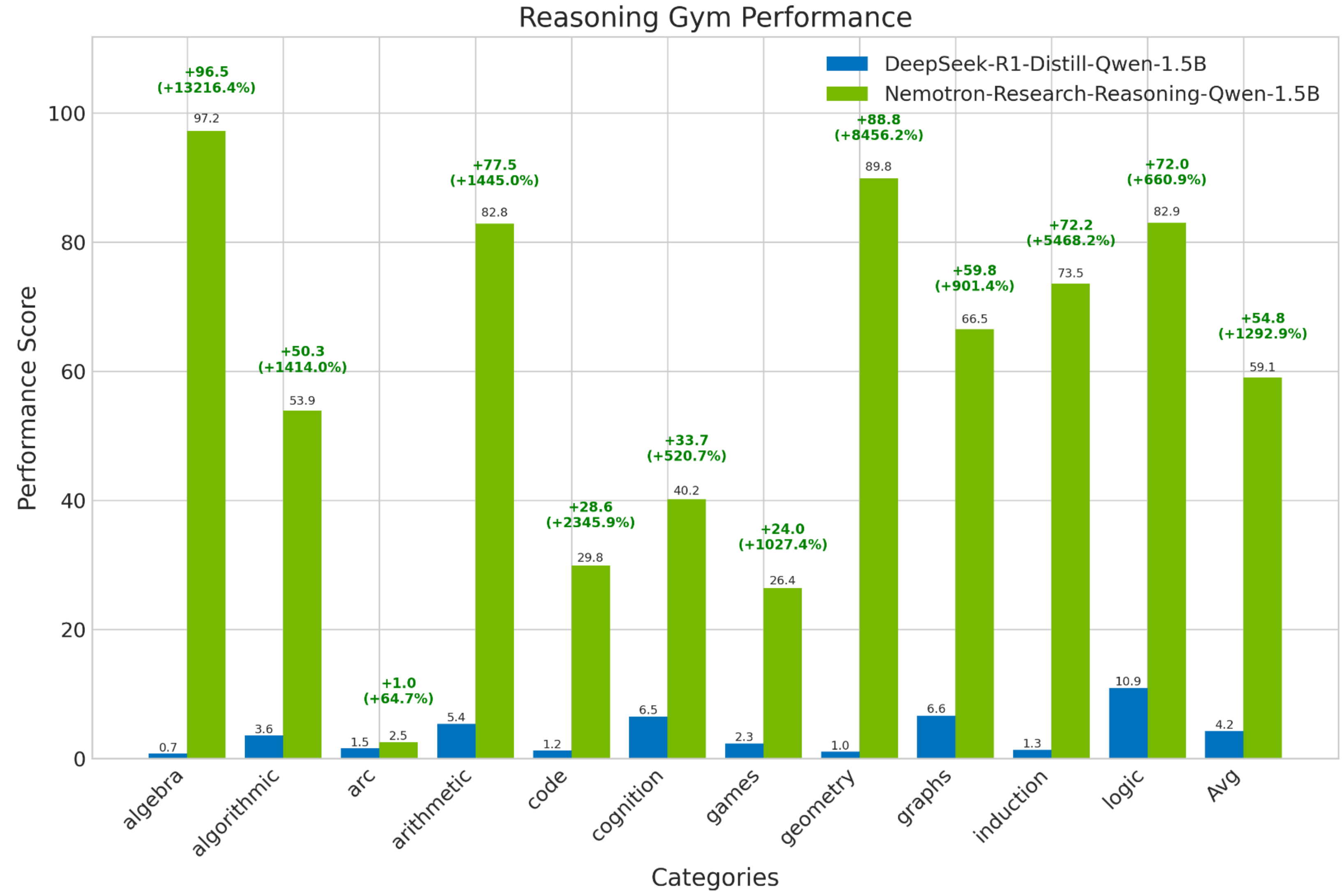
Reasoning Model Training Results

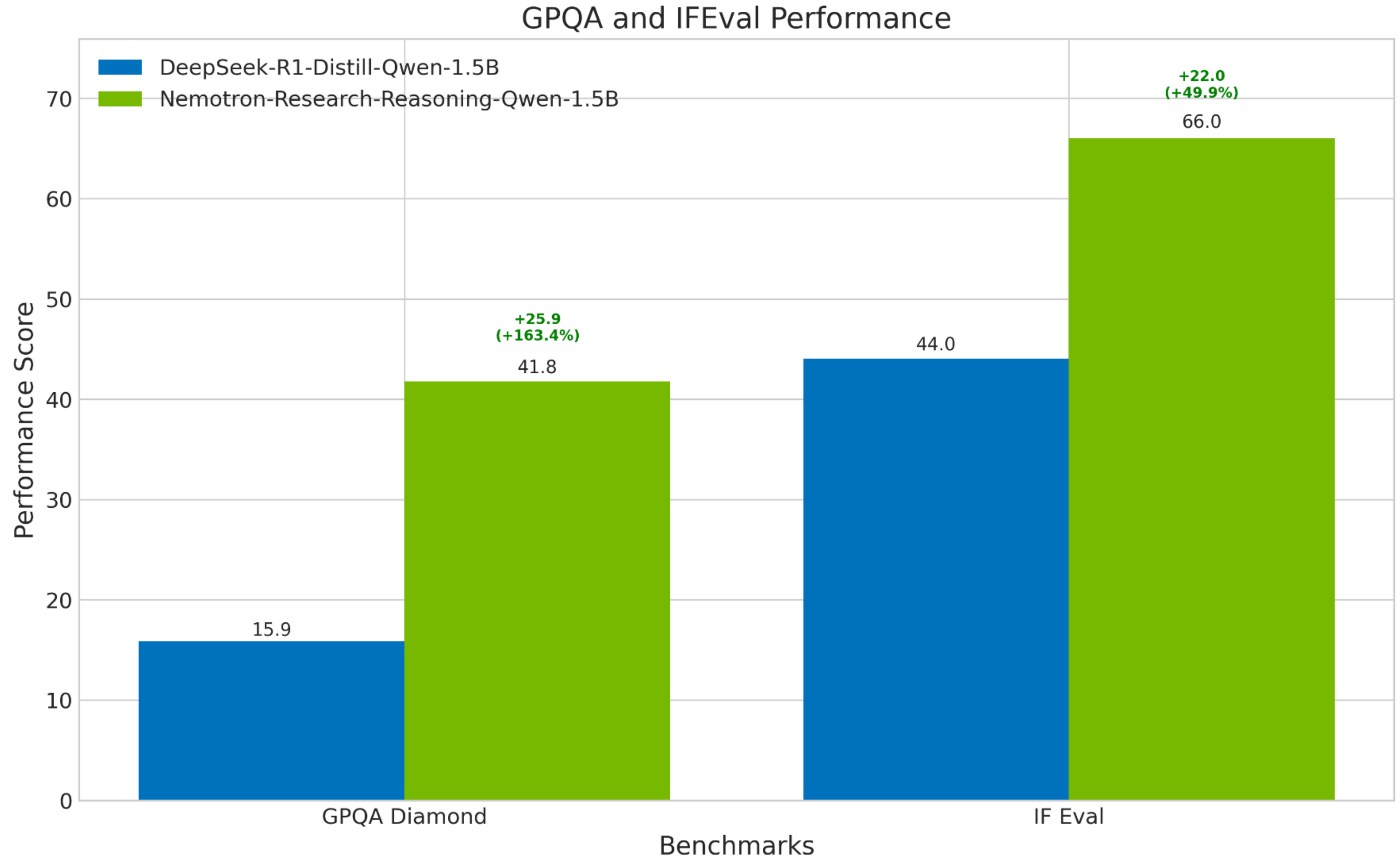
State of Art 1.5B Reasoning Model





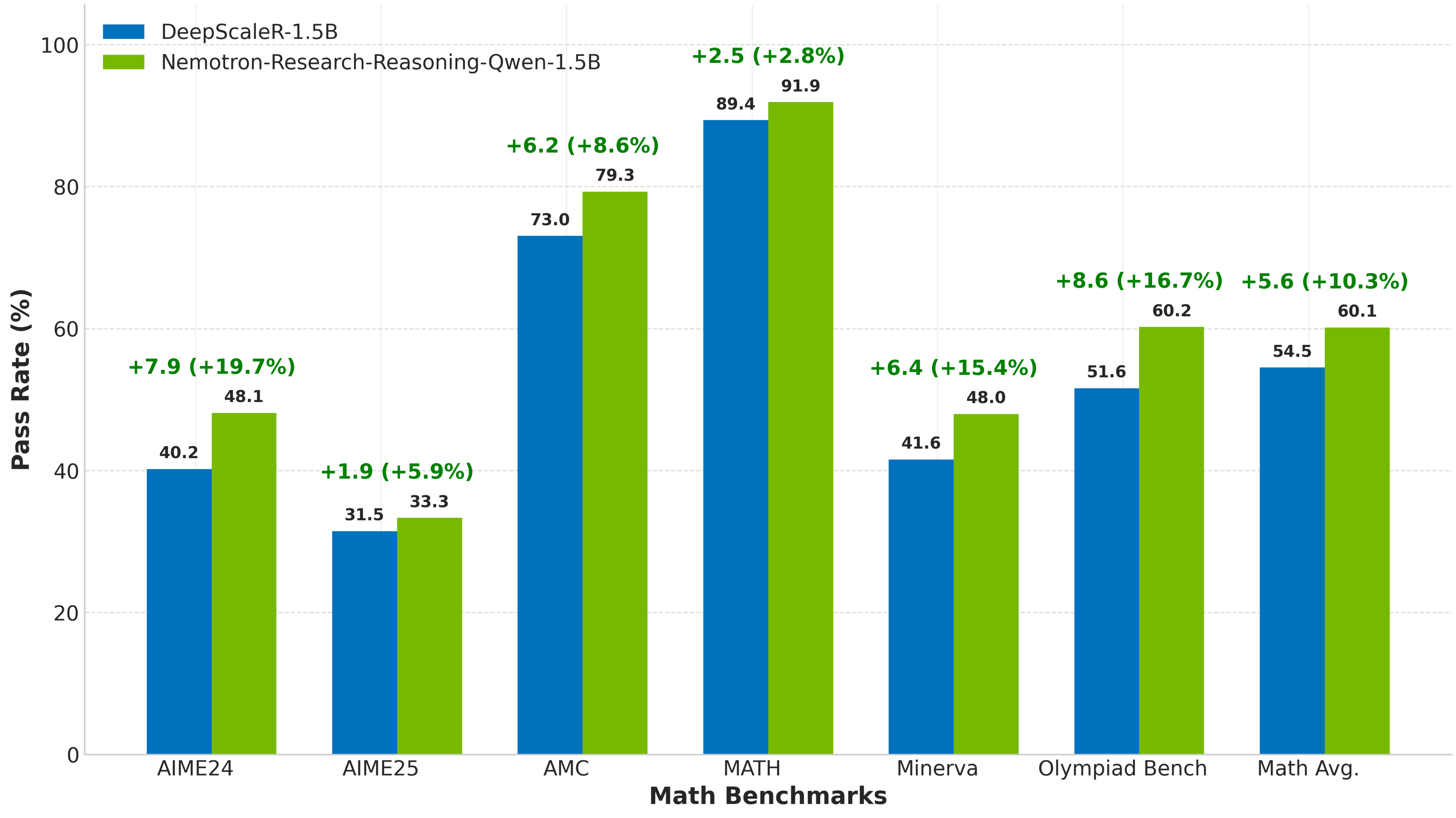


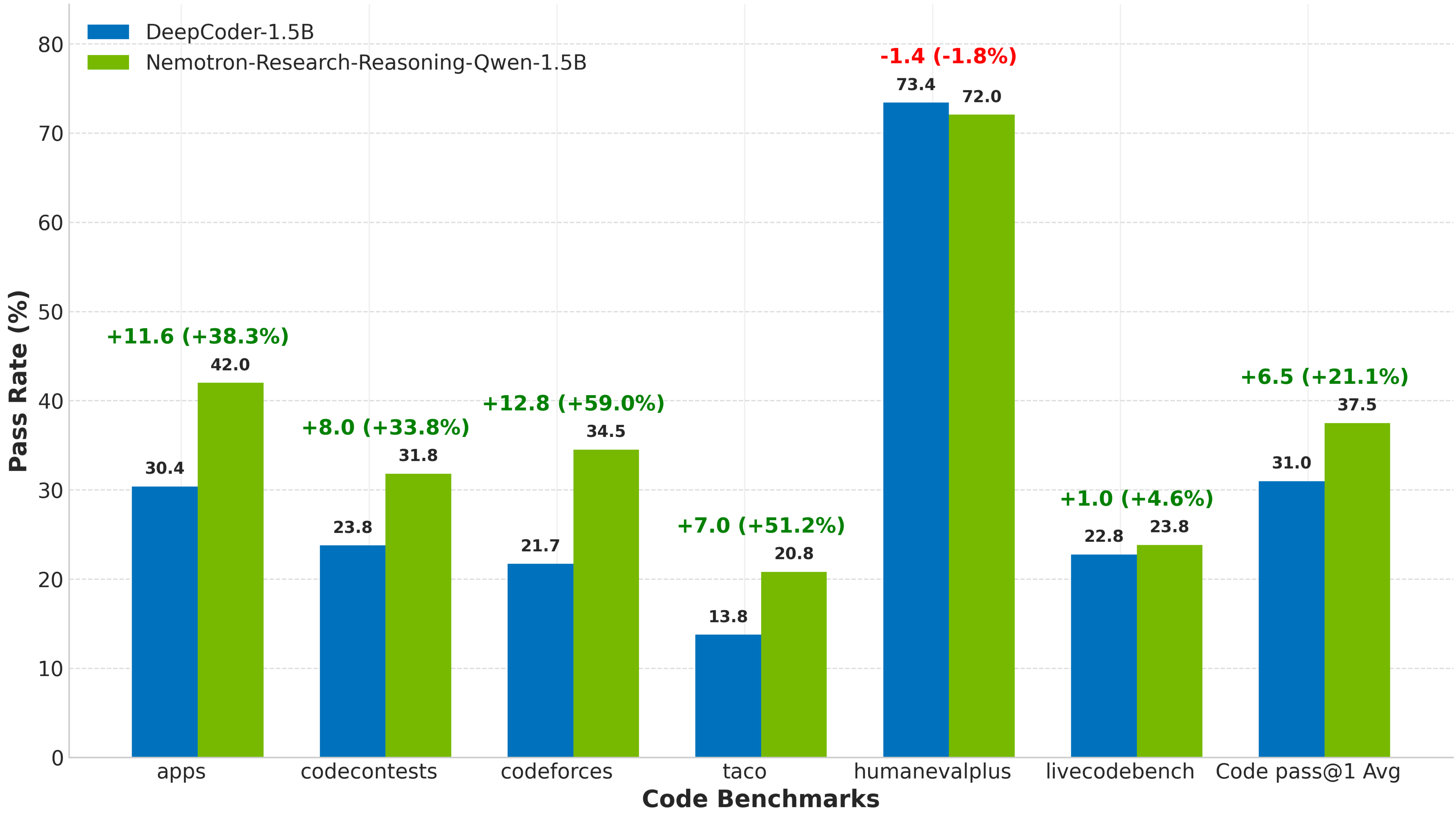




NVIDIA

Better than Specialized Domain Models





Matches DeepSeek-R1-7B Performance

Table 1: Performance (pass@1) comparison for benchmarks across Math domain. The best results are highlighted in **bold**. The results of DeepSeek-R1-Distill-Qwen-7B are marked as gray and are provided as a reference (same in all following tables).

Model	AIME24	AIME25	AMC	Math	Minerva	Olympiad	Avg
DeepSeek-R1-Distill-Qwen-1.5B	28.54	22.71	62.58	82.90	26.38	43.58	44.45
DeepScaleR-1.5B	40.21	31.46	73.04	89.36	41.57	51.63	54.54
DeepSeek-R1-Distill-Qwen-7B	53.54	40.83	82.83	93.68	50.60	57.66	63.19
Nemotron-Research-Reasoning-Qwen-1.5B	**48.13**	**33.33**	**79.29**	**91.89**	**47.98**	**60.22**	**60.14**

Table 2: Performance (pass@1) comparison across benchmarks for Code. We abbreviate benchmarks names for codecontests (cc), codeforces (cf), humanevalplus (human), and livecodebench (LCB).

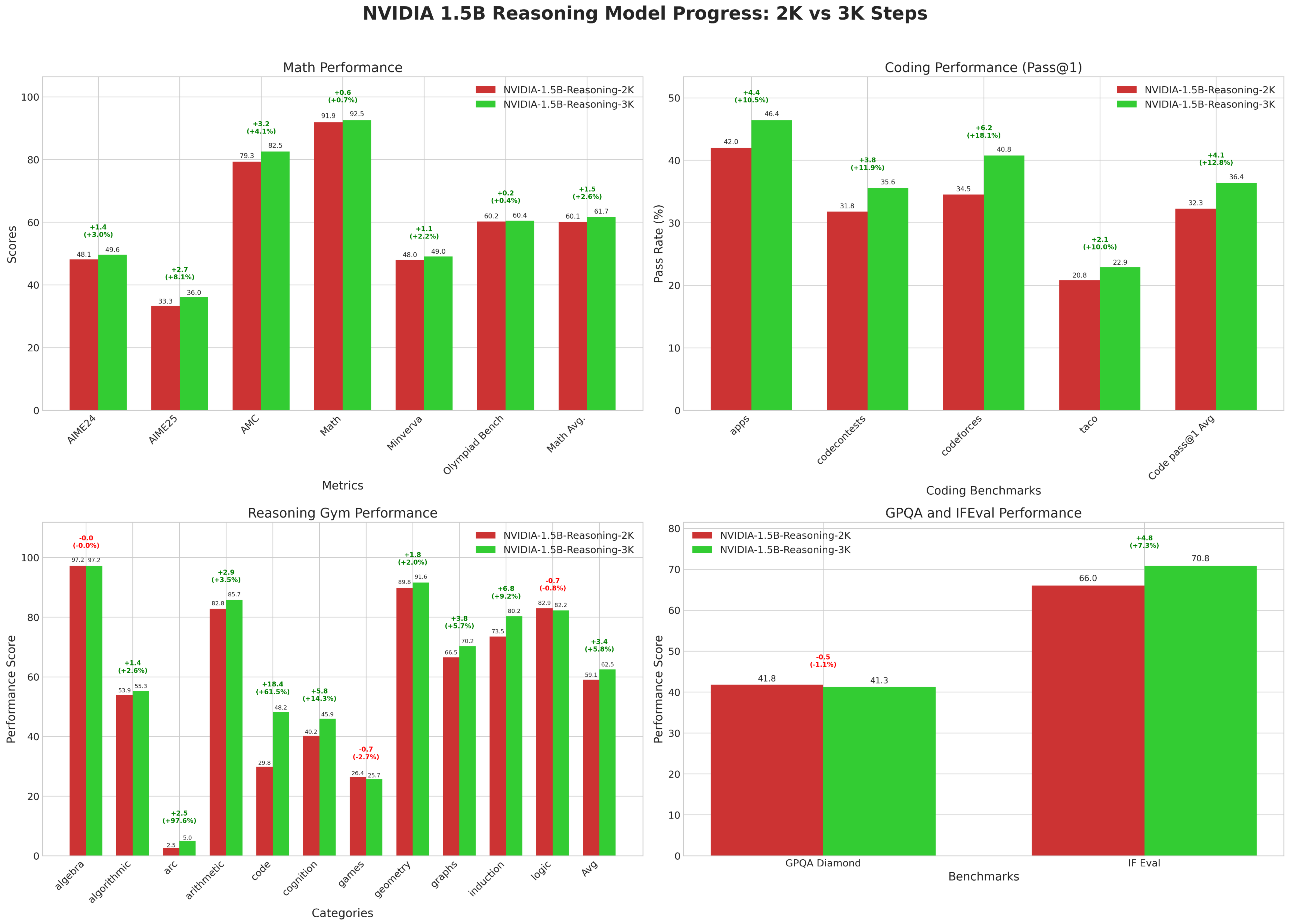
Model	apps	cc	cf	taco	human	LCB	Avg
DeepSeek-R1-Distill-Qwen-1.5B	20.95	16.79	14.13	8.03	61.77	16.80	23.08
DeepCoder-1.5B	30.37	23.76	21.70	13.76	**73.40**	22.76	30.96
DeepSeek-R1-Distill-Qwen-7B	42.08	32.76	33.08	19.08	83.32	38.04	41.39
Nemotron-Research-Reasoning-Qwen-1.5B	**41.99**	**31.80**	**34.50**	**20.81**	72.05	**23.81**	**37.49**

Table 3: Performance comparison on STEM reasoning (GPQA Diamond), instruction following (IFEval), and logic puzzles (Reasoning Gym) tasks. We also present results on OOD tasks: *acre*, *boxnet*, and *game_of_life_halting* (game).

Model	GPQA	IFEval	Reasoning	acre	boxnet	game
DeepSeek-R1-Distill-Qwen-1.5B	15.86	44.05	4.24	5.99	0.00	3.49
DeepSeek-R1-Distill-Qwen-7B	35.44	58.01	28.55	20.21	1.71	12.94
Nemotron-Research-Reasoning-Qwen-1.5B	**41.78**	**66.02**	**59.06**	**58.57**	**7.91**	**52.29**

Keep training our 1.5B model

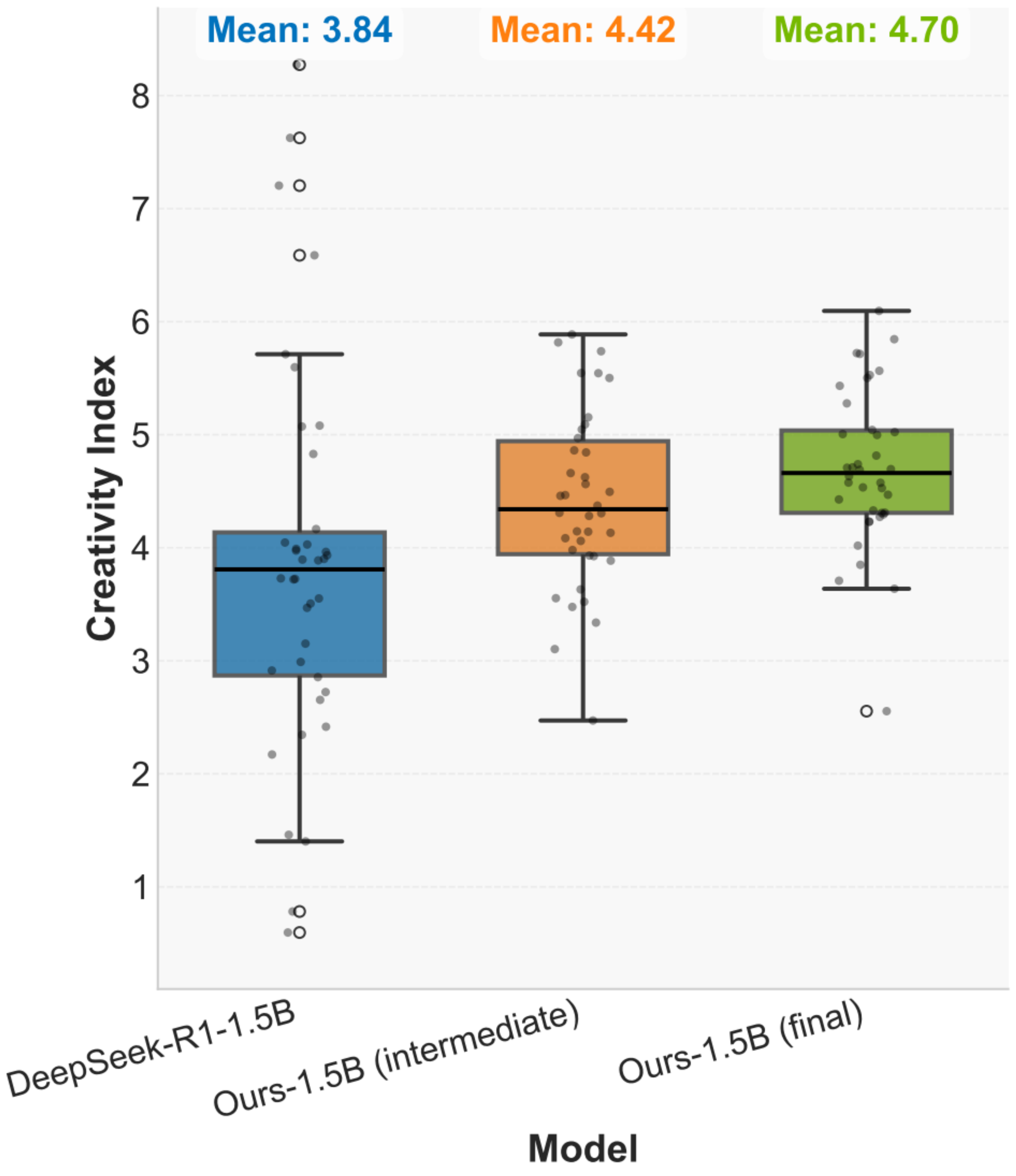
3K steps

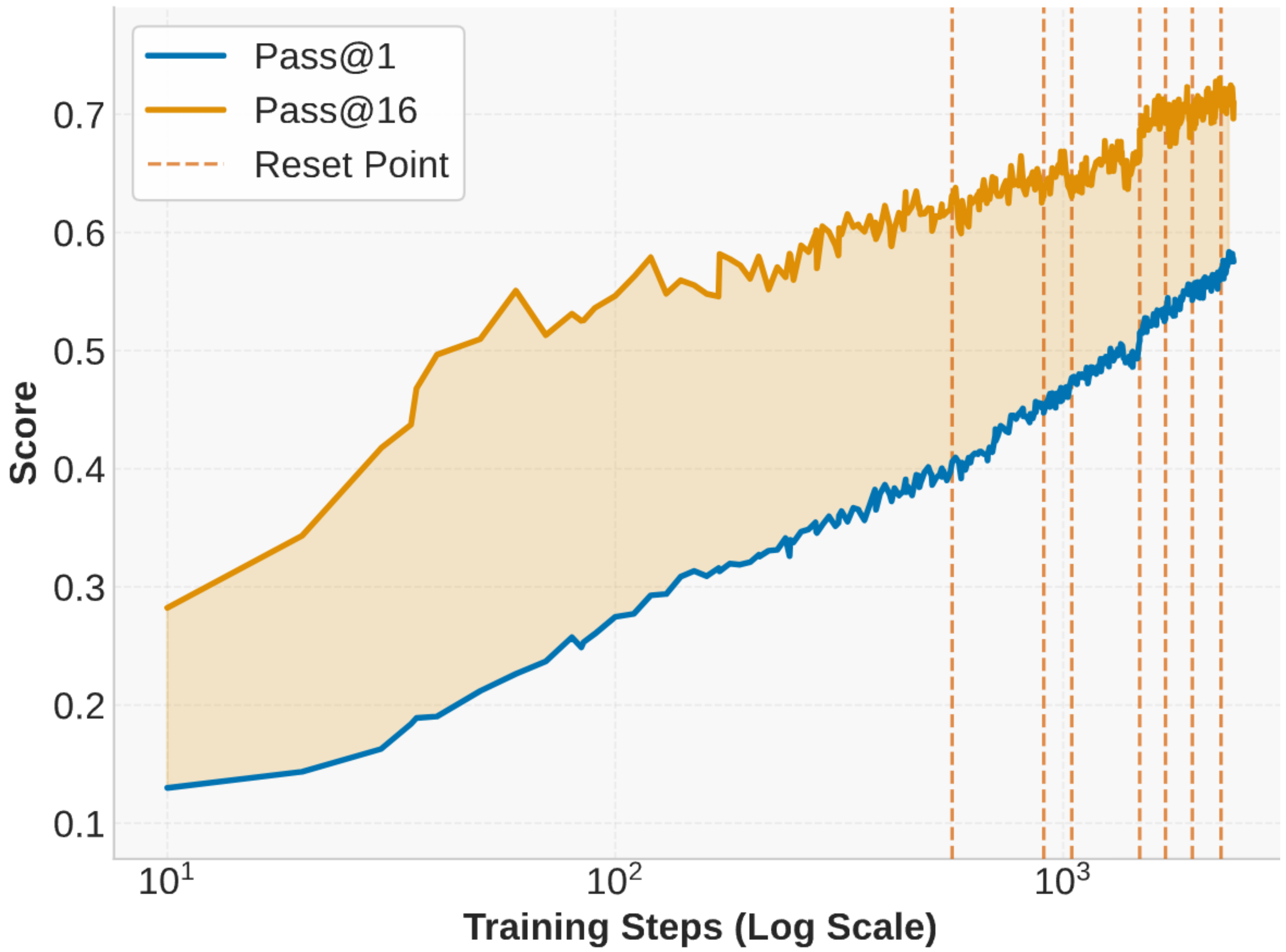


Reasoning Capabilities Expand With Continued Training.

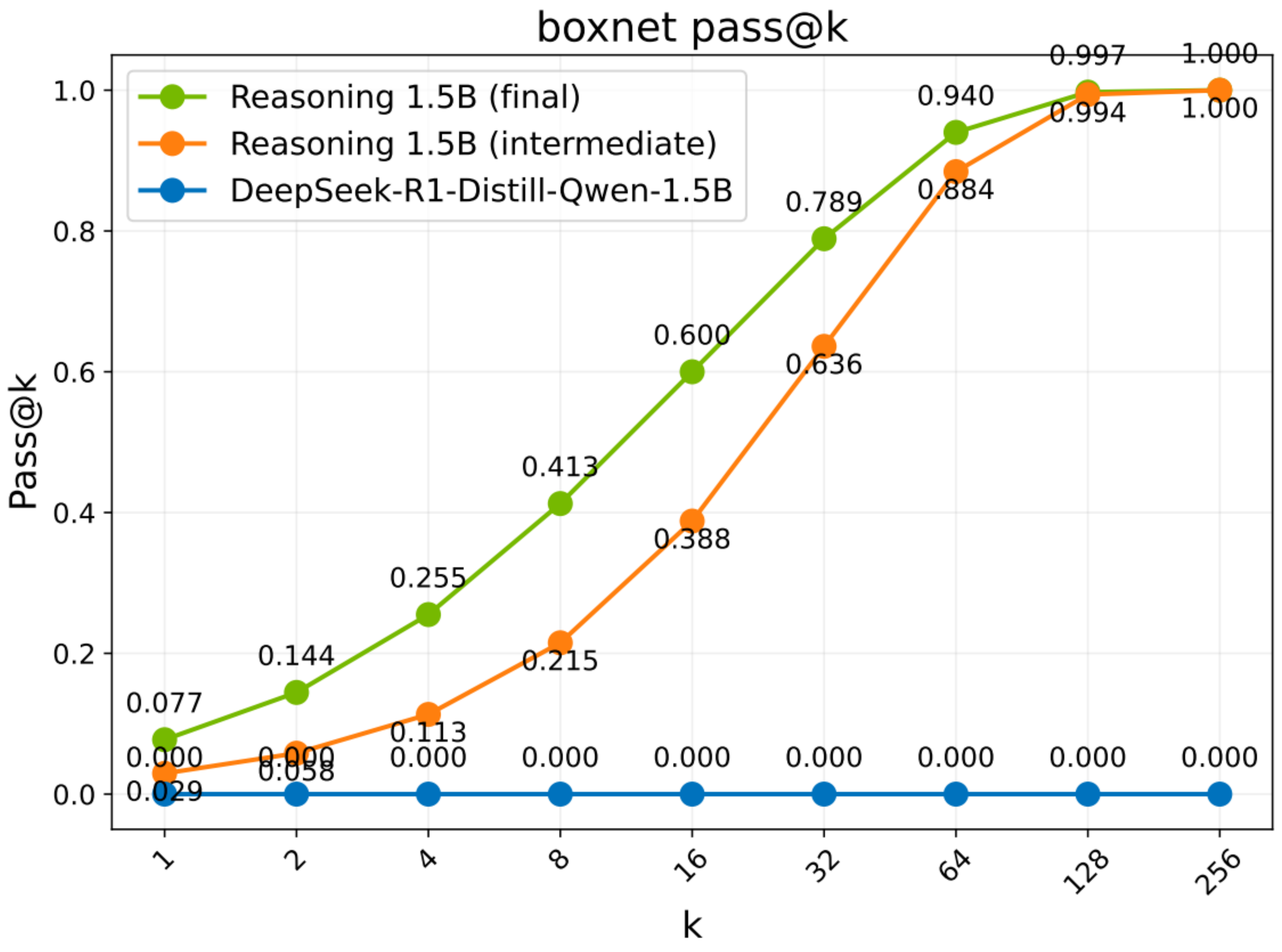
ProRL makes a difference

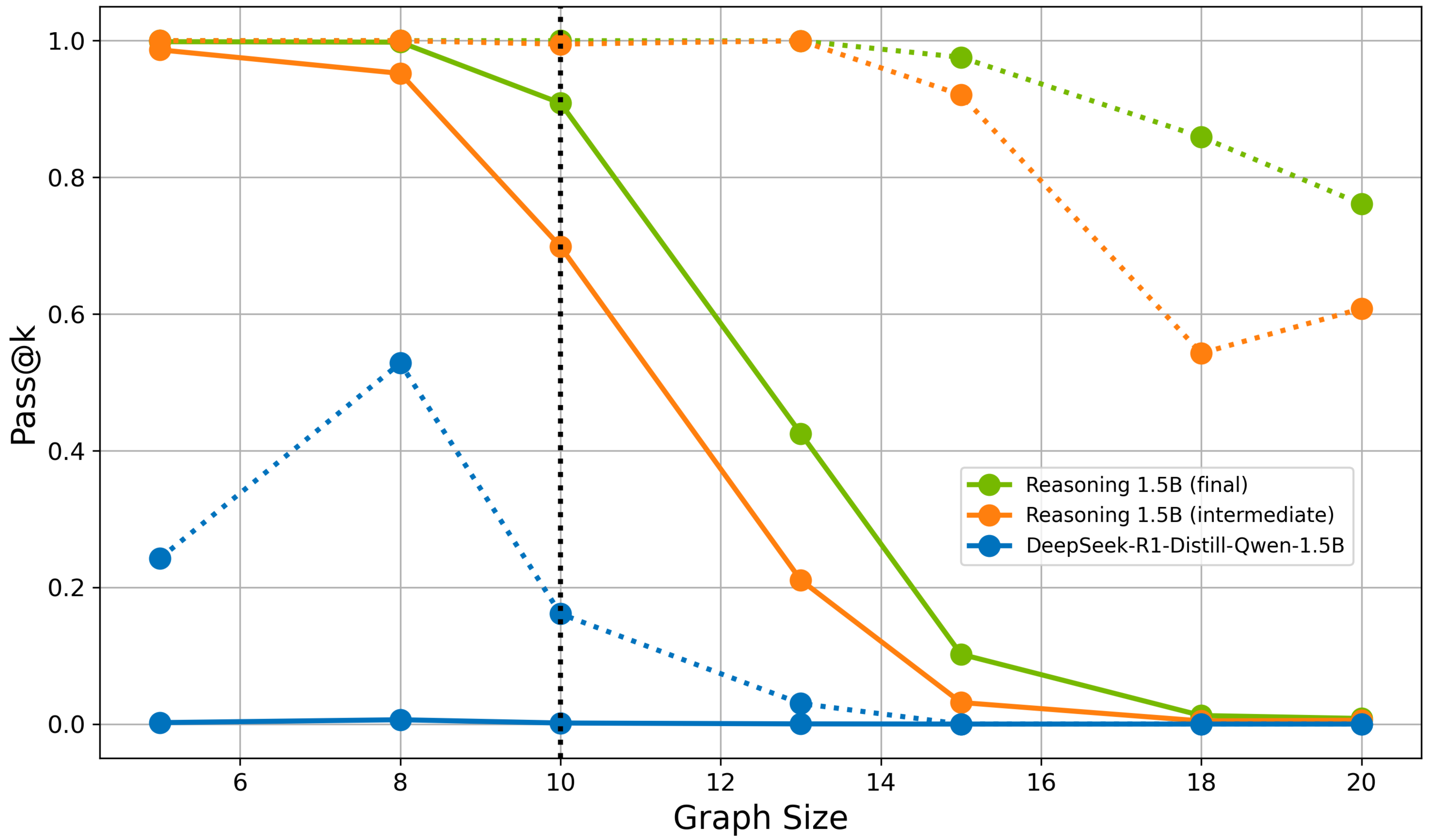
Continued Training Improves Creativity Index and Pass@K



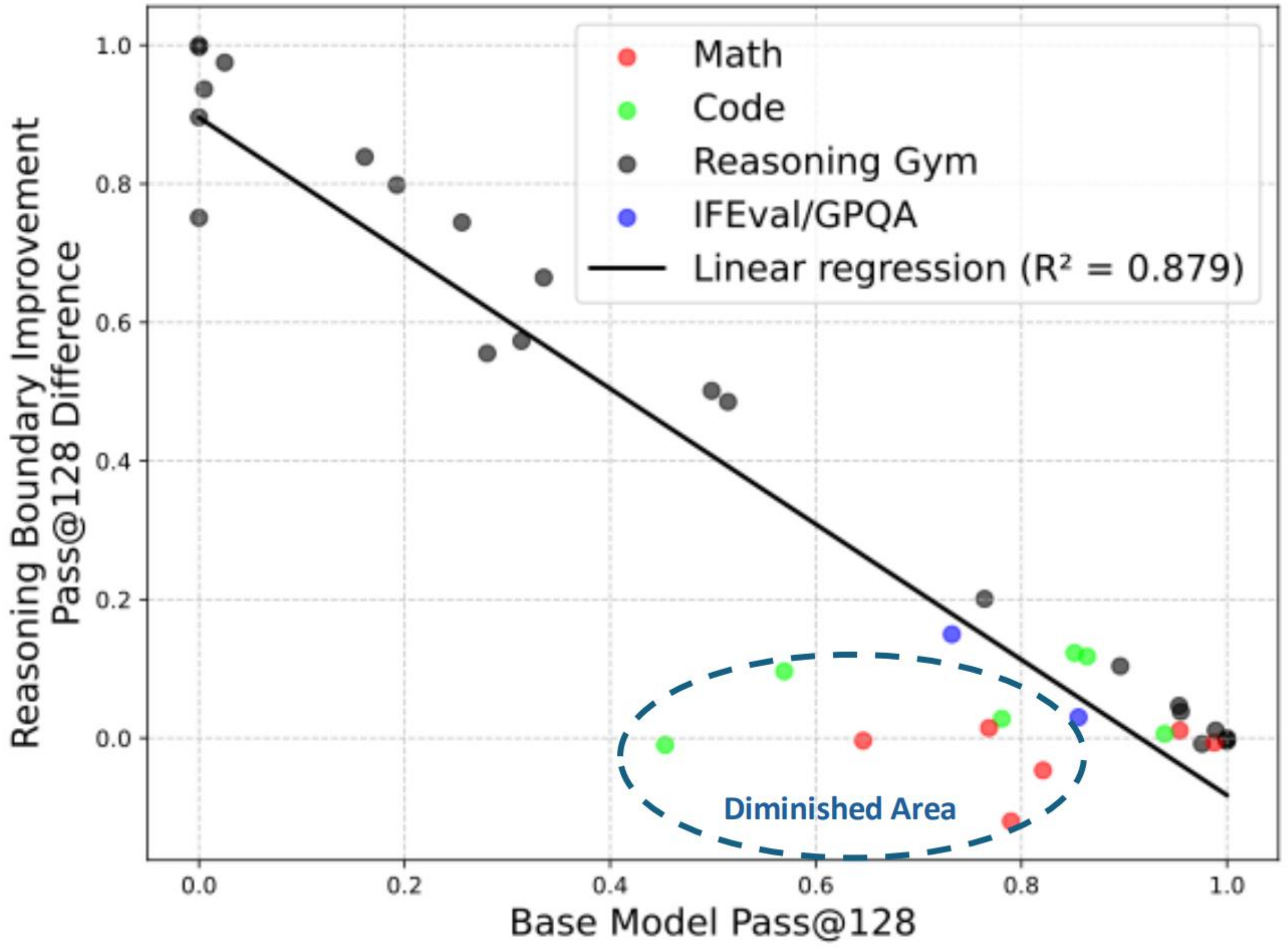


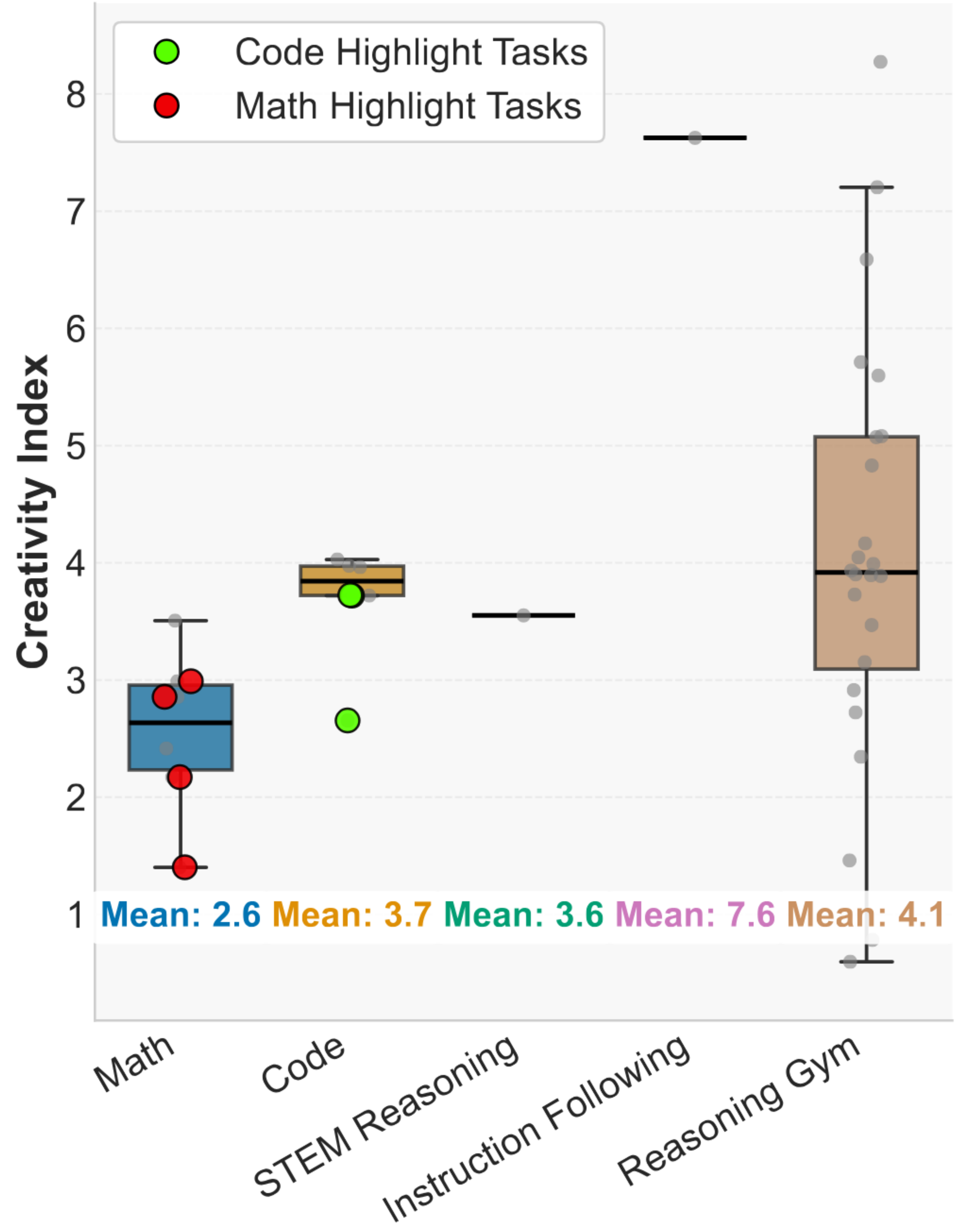
Out of Distribution For Different Task and Difficulty Levels





Explain the Reasoning Boundary Improvement





Summary

- Our findings suggest that prolonged reinforcement learning (ProRL) training significantly expands the reasoning capabilities of base models.
- The more compute ProRL uses, the more creative solutions it discovers.
- Previous claims about temperature distillation effect were likely due to:
  - A narrow focus on overtrained mathematical tasks
  - Insufficient RL training steps
- In contrast, our approach emphasizes scaling reasoning model training through a wide range of diverse and novel tasks.
- To fully realize the benefits of ProRL, it is essential to maintain stable entropy and periodically reset the reference model to eliminate performance boundaries.
- Using ProRL, we successfully trained a state-of-the-art 1.5B parameter reasoning model.

NVIDIA