The Art of Artificial Reasoning for (Small) Language Models

Yejin Choi

Stanford & NVIDIA 💚

The background art created by Jack Hessel, using DALL.E of OpenAI

Mission Impossible: Democratizing Generative AI by Transcending Scaling Laws

because (current) scaling laws demand extreme-scale compute that only a few can afford

Isn't that just impossible?

😱

A case in point…

Circa 2023

How can Indian startups create foundation models for India?

Rajan Anandan

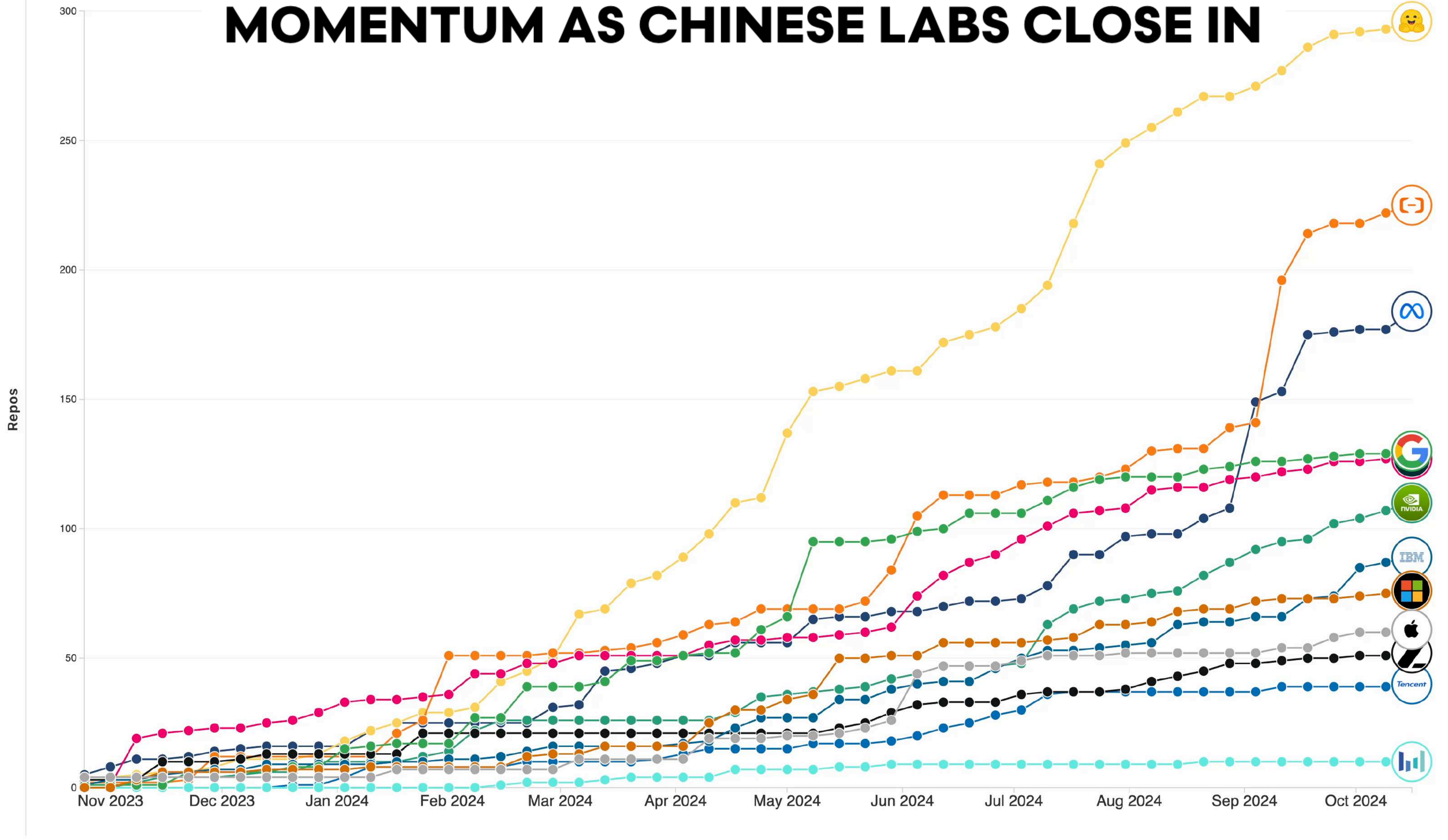
Sam Altman

It's hopeless to compete with OpenAI

https://aiworld.eu/story/nvidia-leads-open-source-ai-momentum-as-chinese-labs-close-in

AI WORLD

NVIDIA LEADS OPEN SOURCE AI MOMENTUM AS CHINESE LABS CLOSE IN



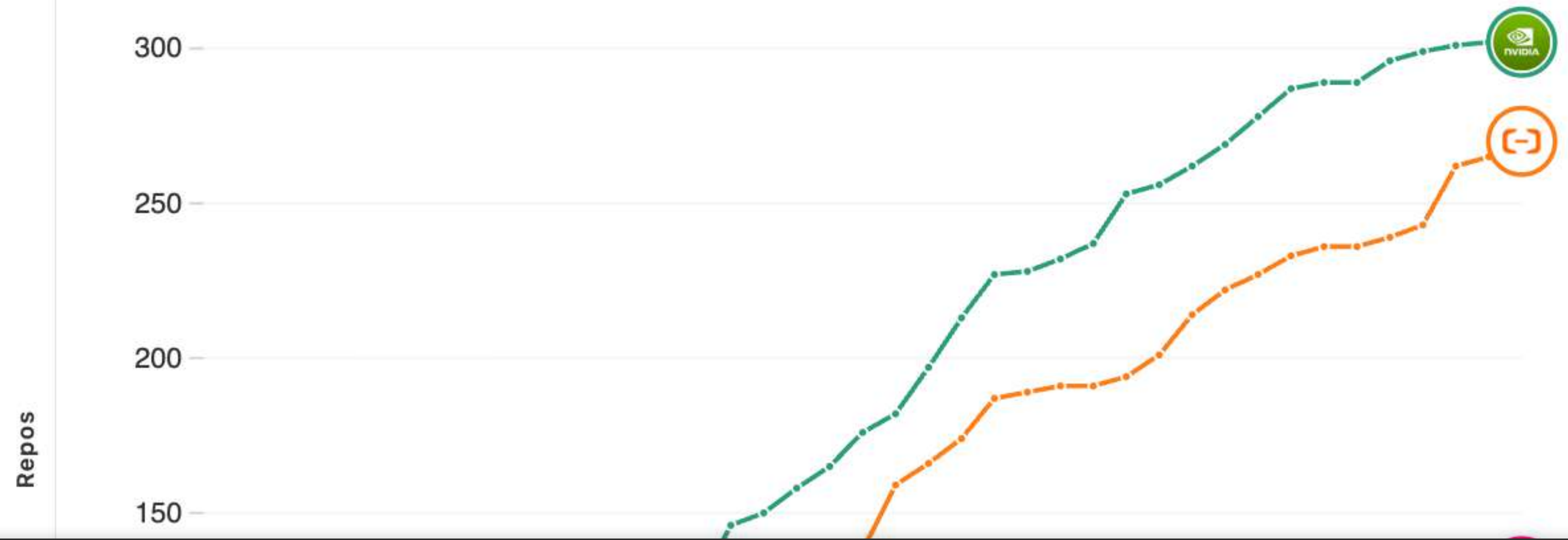
October 23, 2024

* Considering repos above 10 space likes and 500 downloads

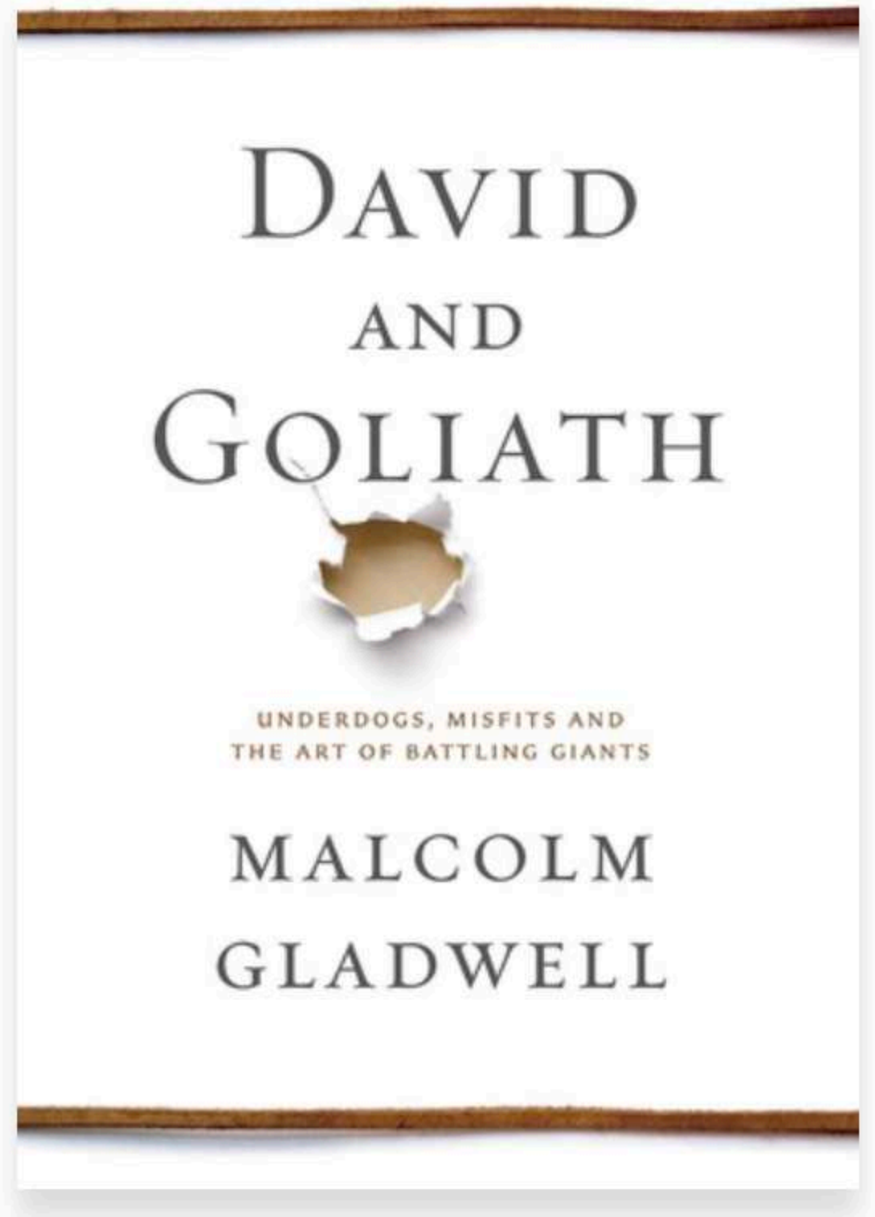
AI WORLD

https://aiworld.eu/story/nvidia-leads-open-source-ai-momentum-as-chinese-labs-close-in

NVIDIA LEADS OPEN SOURCE AI MOMENTUM AS CHINESE LABS CLOSE IN



1. Small models are in high demand, with daily downloads hitting tens of millions (!!! 🔥) for top repositories
2. China's contribution to the open source ecosystem has skyrocketed, with Alibaba now surpassing Hugging Face 😮
3. NVIDIA has emerged as a rising star ⭐ (or a risen star? 😎), for open-source contributions 💚



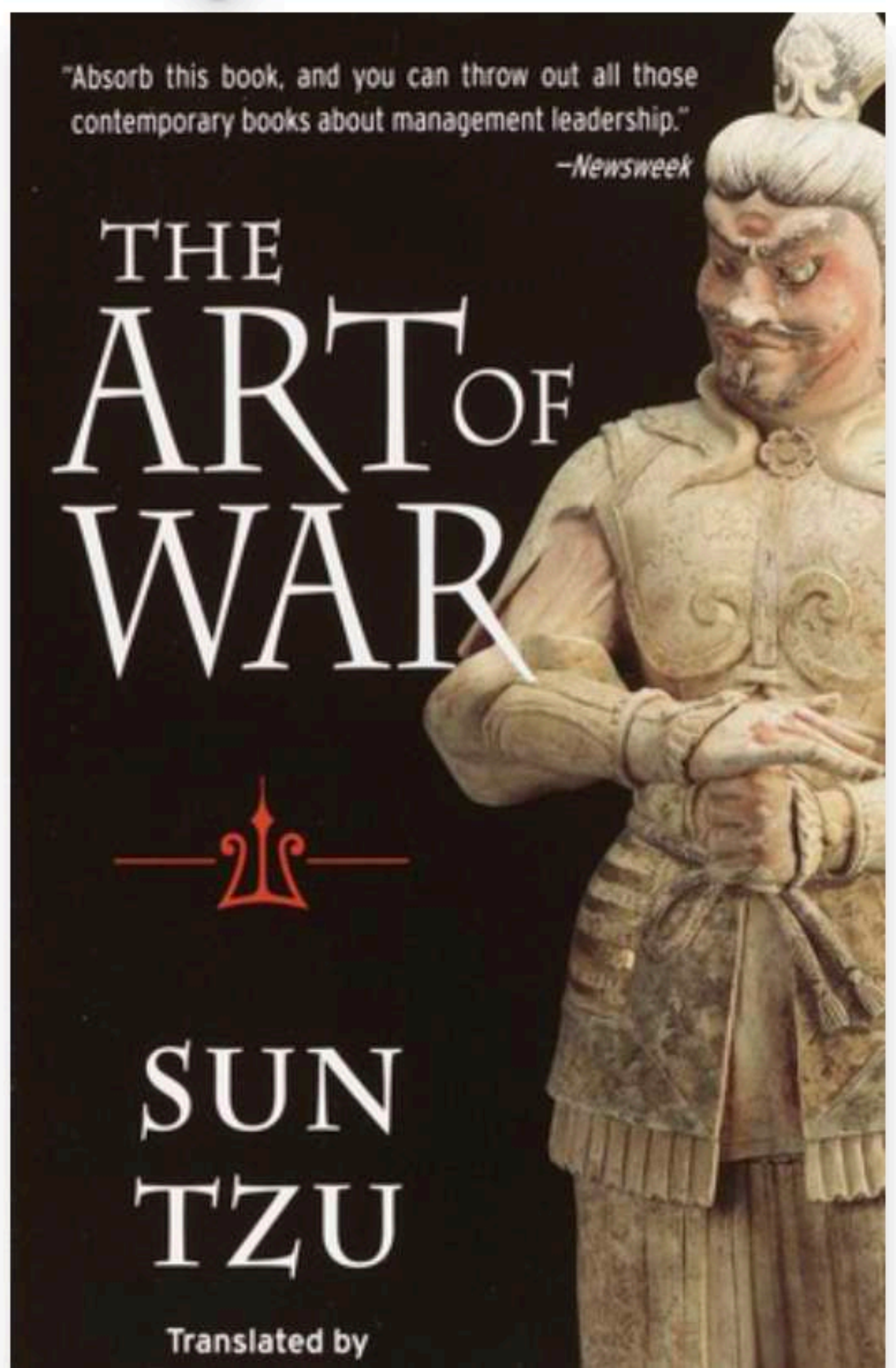
DAVID
AND
GOLIATH

UNDERDOGS, MISFITS AND
THE ART OF BATTLING GIANTS

MALCOLM
GLADWELL

David and Goliath: Underdogs, Misfits, and the Art of Battling Giants

by Malcolm Gladwell


"Absorb this book, and you can throw out all those contemporary books about management leadership."
—Newsweek
THE
ART OF
WAR
SUN
TZU
Translated by

The Art of War

by Sun Tzu, Thomas Cleary
Pulat Otkan (Translator)

David vs. Goliath:
the Art of Scaling Intelligence
in the Era of Extreme-Scale Neural Models

Three components to innovate:

Unconventional data 🔥

Unconventional algorithms 🚀

Unconventional collaboration 🌏

The Era of Brute-Force Scaling is Over

The Era of Smart Scaling Begins

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

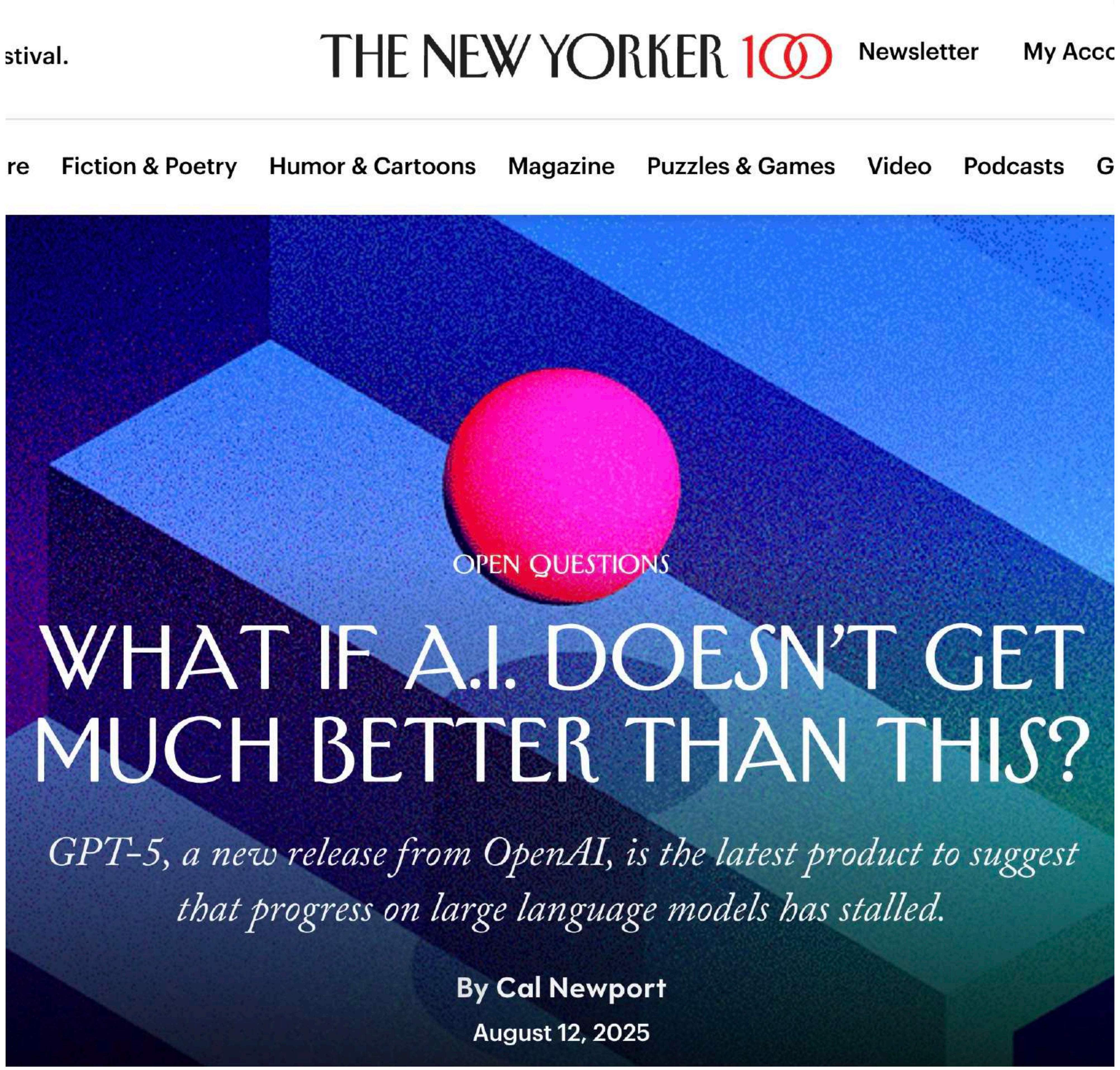
Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

-- Ilya Sutskever from his test time award talk at NeurIPS 2024

The Era of Brute-Force Scaling is Over

The Era of Smart Scaling Begins



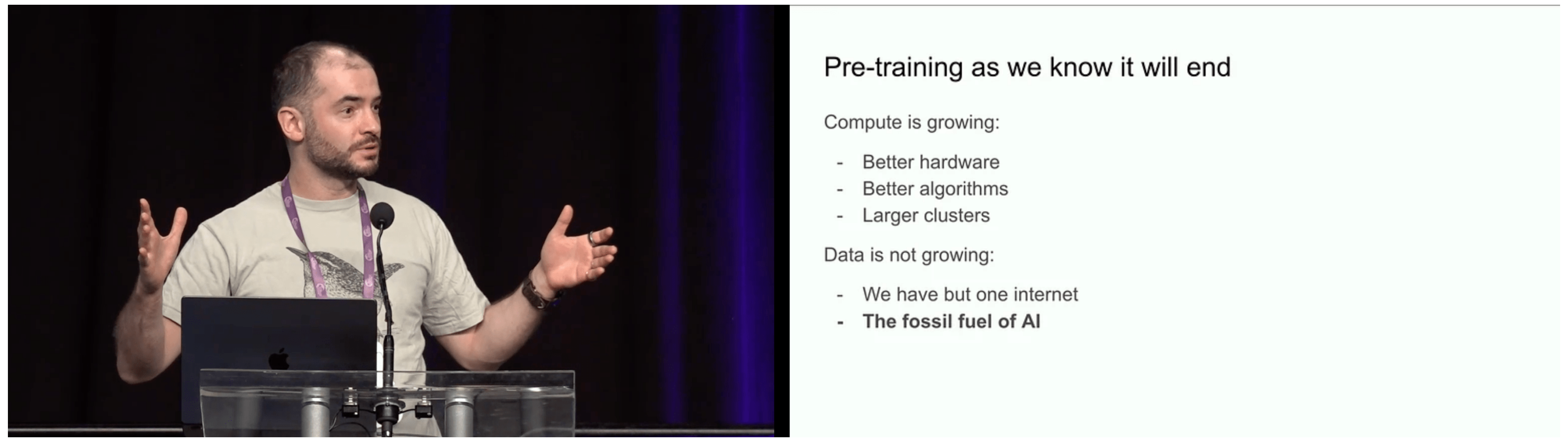
OPEN QUESTIONS

WHAT IF A.I. DOESN'T GET MUCH BETTER THAN THIS?

GPT-5, a new release from OpenAI, is the latest product to suggest that progress on large language models has stalled.

By Cal Newport

August 12, 2025



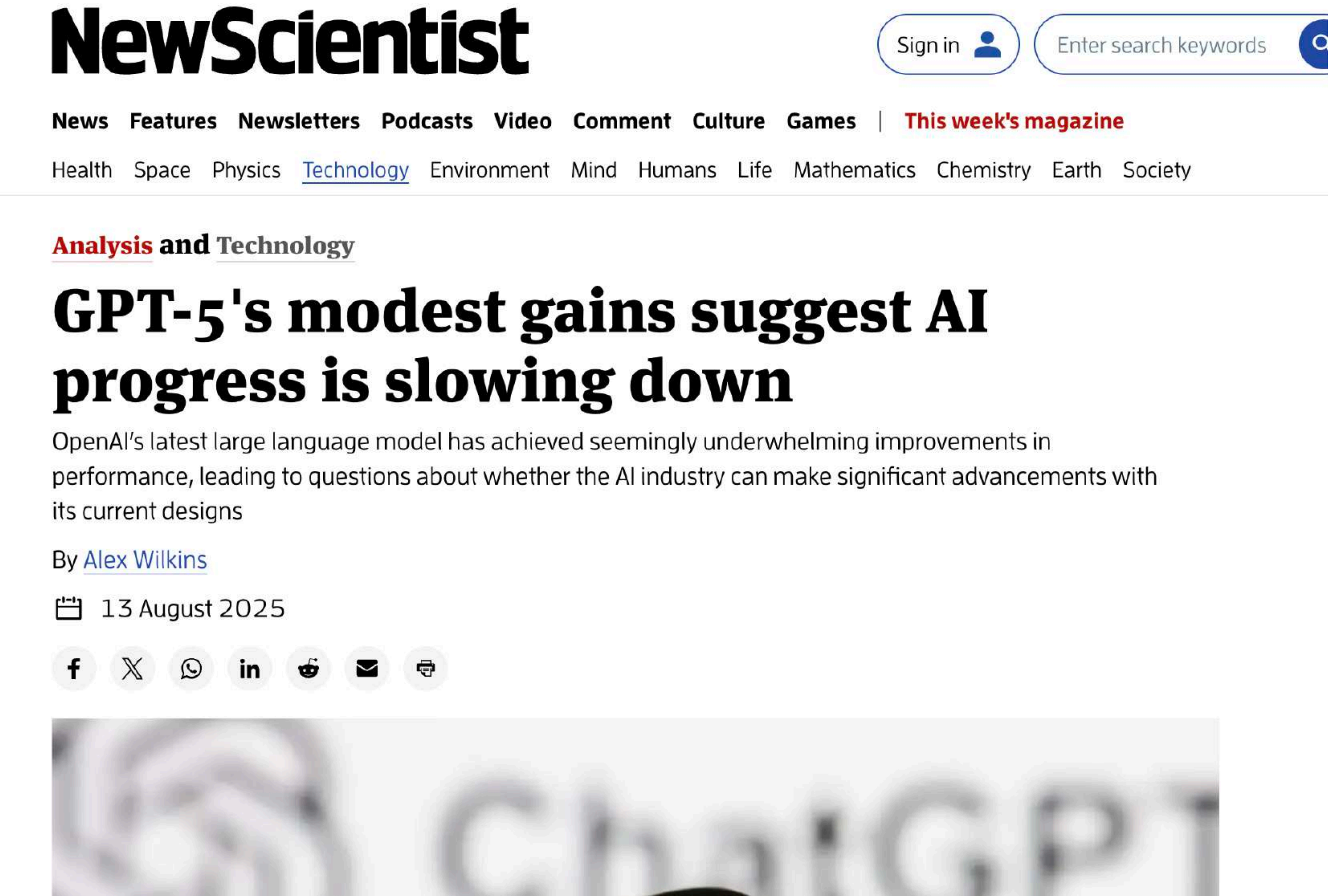
Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**



Analysis and Technology

GPT-5's modest gains suggest AI progress is slowing down

OpenAI's latest large language model has achieved seemingly underwhelming improvements in performance, leading to questions about whether the AI industry can make significant advancements with its current designs

By Alex Wilkins

13 August 2025

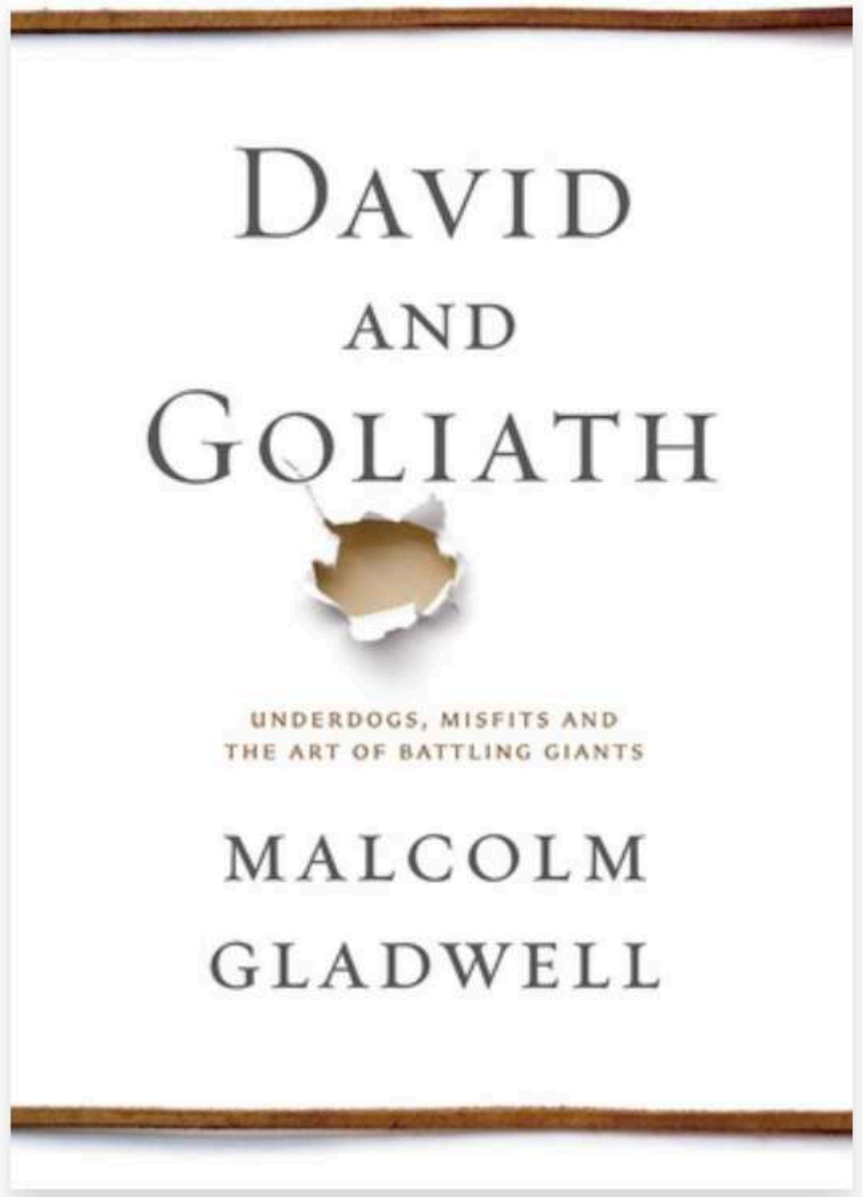
The Era of Brute-Force Scaling is Over

The Era of Smart Scaling Begins

Three broad ways to cope with data saturation:

1. Learn better & faster with limited data
   - Alternative architectures
   - Alternative training recipes
2. Synthesize new data
   - Algorithms to generate the outer space of the internet data
3. Reason beyond what is in the data
   - Test-time reasoning algorithms
   - Test-time training algorithms

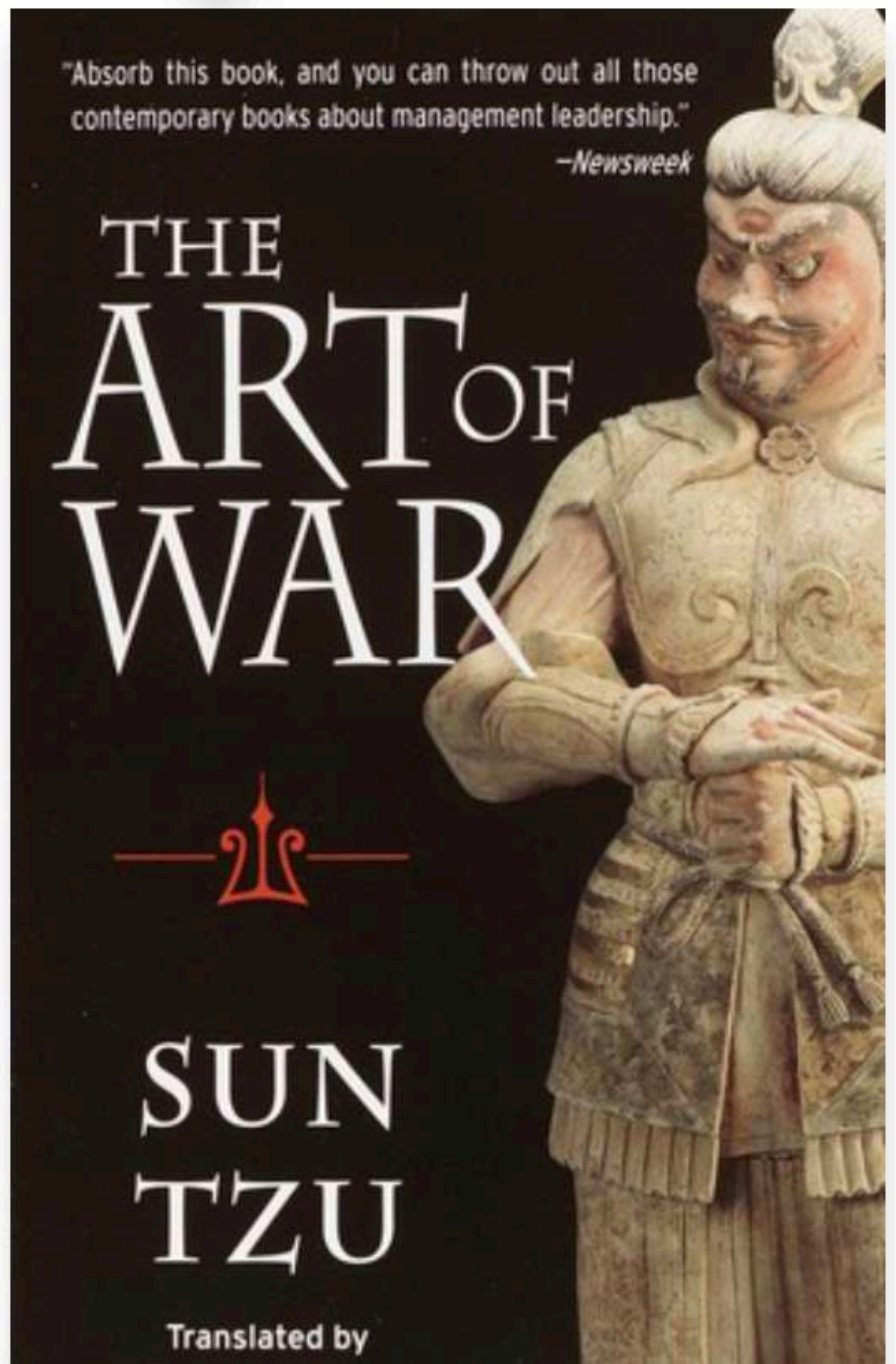
=> Scaling of “intelligence” continues



David and Goliath: Underdogs, Misfits, and the Art of Battling Giants

by Malcolm Gladwell

David vs. Goliath:
the Art of Scaling Intelligence
in the Era of Extreme-Scale Neural Models


"Absorb this book, and you can throw out all those contemporary books about management leadership." —Newsweek
THE
ART OF
WAR
SUN
TZU
Translated by

The Art of War

by Sun Tzu, Thomas Cleary
Pulat Otkan (Translator)

Three components to innovate:
Unconventional data 🔥
Unconventional algorithms 🚀
Unconventional collaboration 🌏

In this talk:

===

ProRL: Prolonged RL

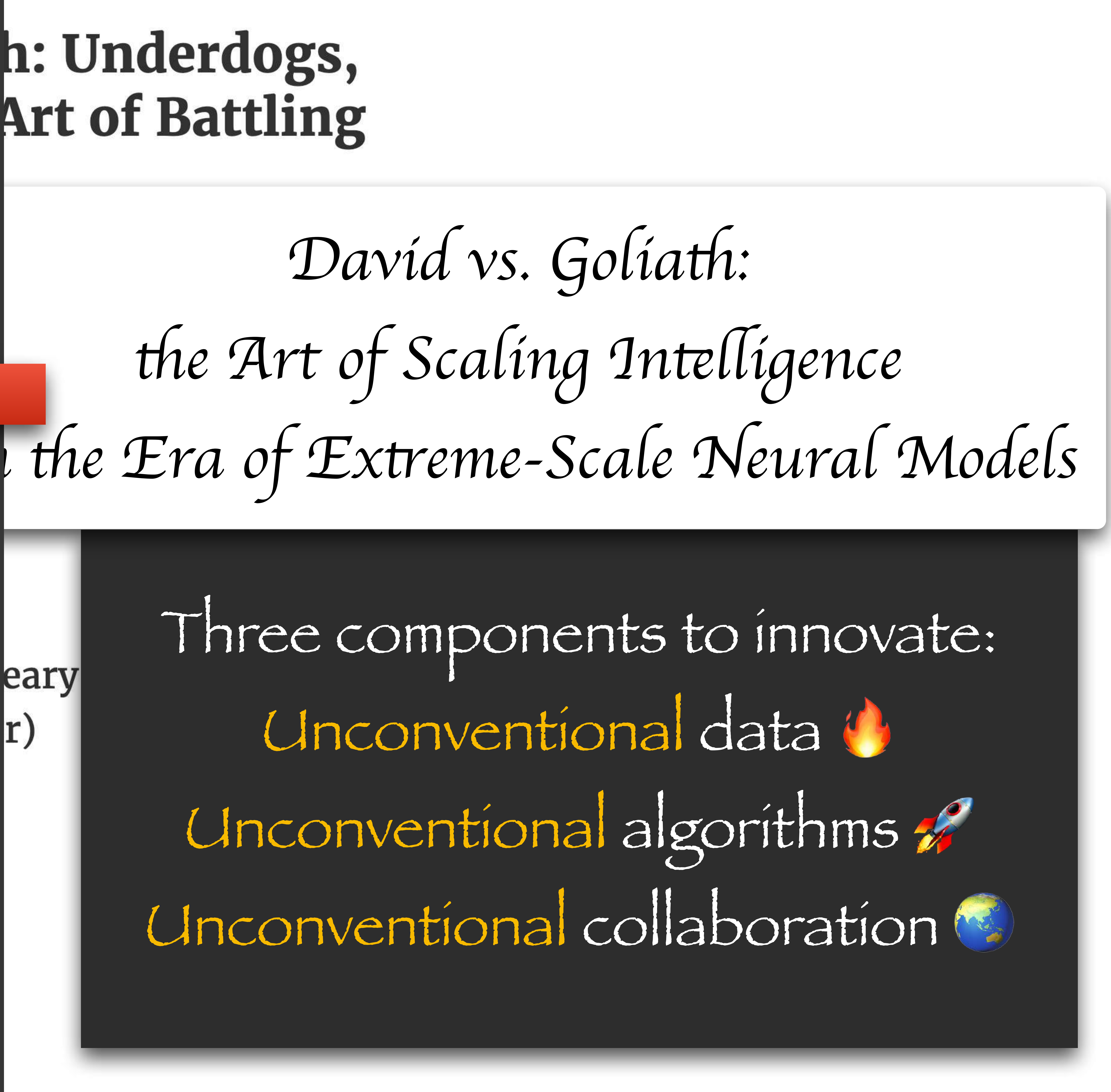
Prismatic Synthesis

RL as Pretraining

===

“Smaller but Better”

“Algorithms for the Win”



h: Underdogs,
Art of Battling

eary
r)

David vs. Goliath:
the Art of Scaling Intelligence
the Era of Extreme-Scale Neural Models

Three components to innovate:

Unconventional data 🔥

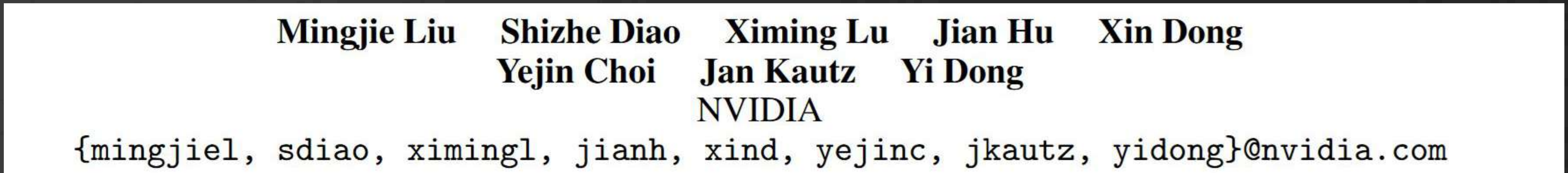
Unconventional algorithms 🚀

Unconventional collaboration 🌏

ProRL

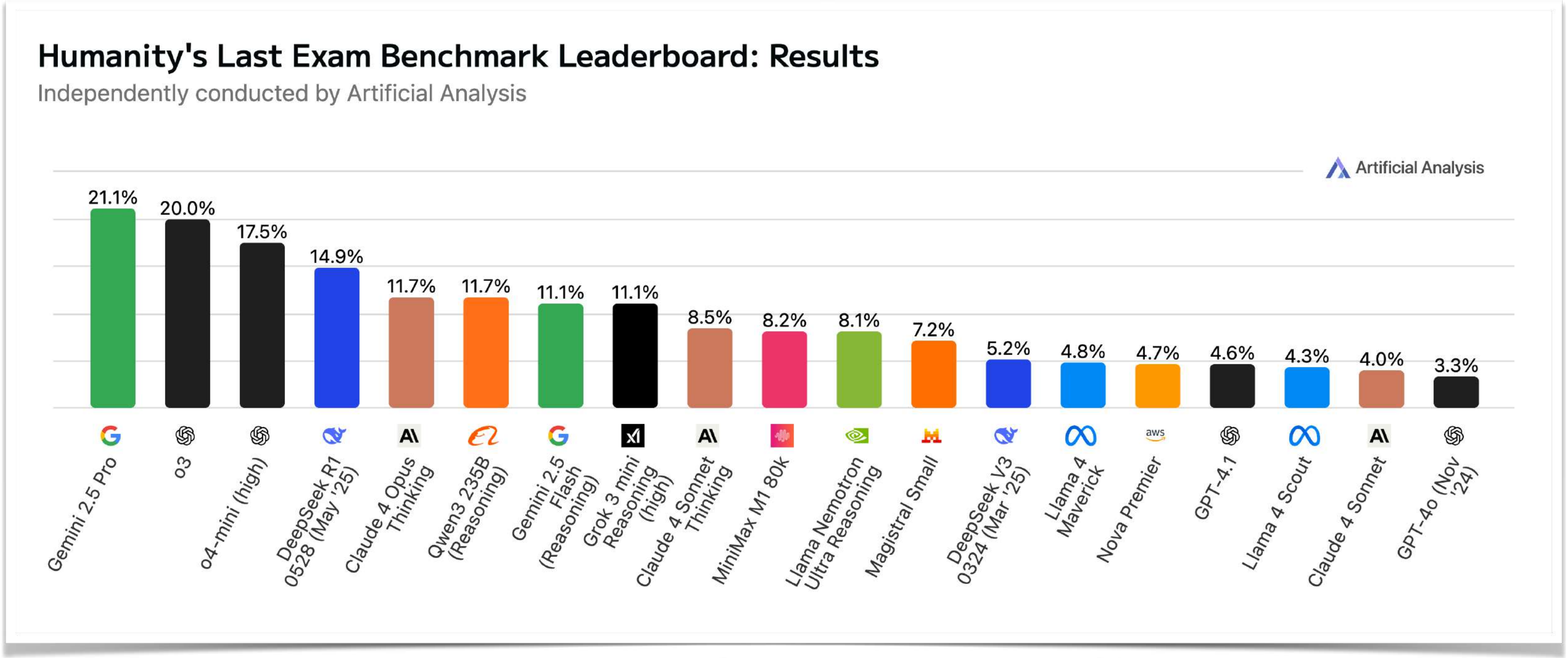
— NeurIPS 2025 —

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: The Rise of LRMs (as opposed to LLMs)

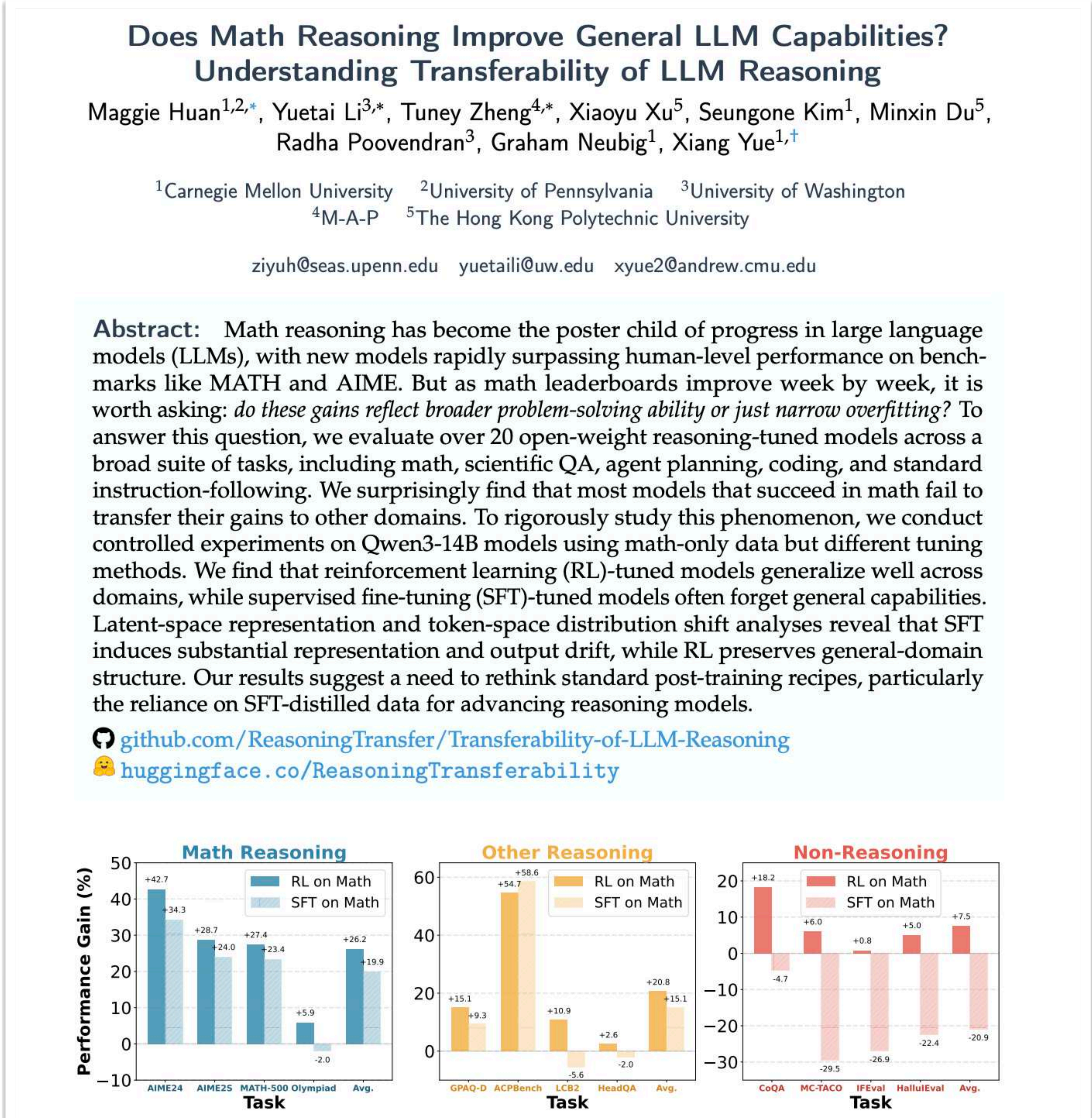


Humanity's Last Exam Benchmark Leaderboard: Results

Independently conducted by Artificial Analysis

- **Long thought** (chain-of-thought)
- The power of **Reinforcement Learning**
- Shift from **Imitation Learning** to **Exploration Learning**

2025: The Rise of LRMs (as opposed to LLMs)



Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

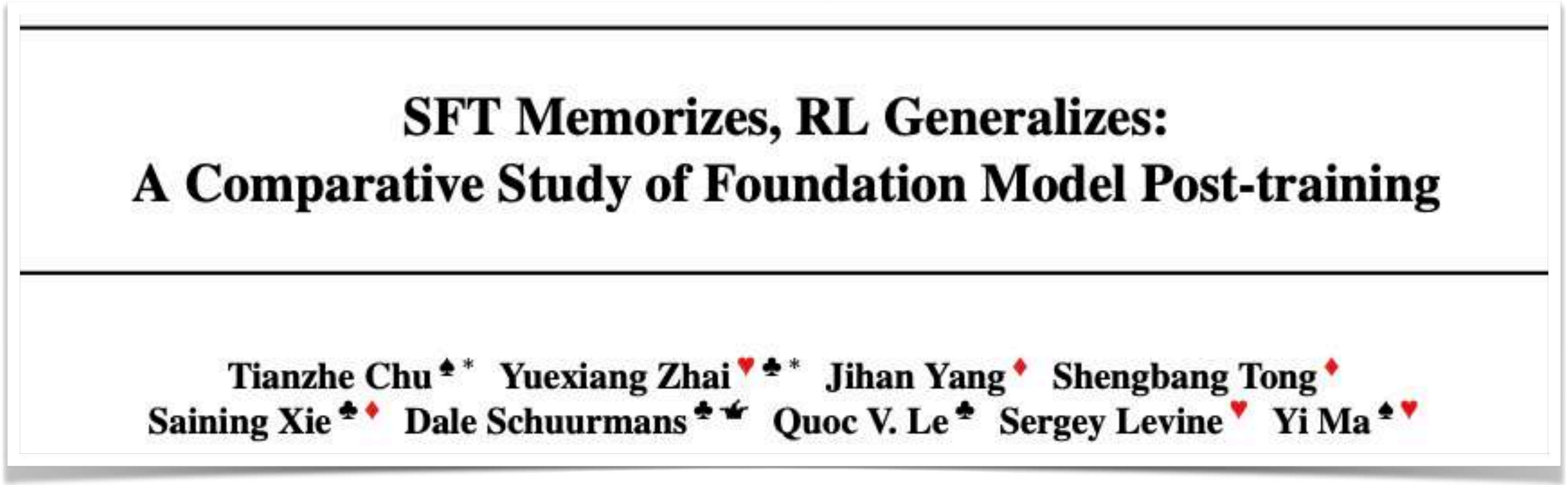
Maggie Huan[1,2,*], Yuetai Li[3,*], Tuney Zheng[4,*], Xiaoyu Xu[5], Seungone Kim[1], Minxin Du[5], Radha Poovendran[3], Graham Neubig[1], Xiang Yue[1,†]

[1]Carnegie Mellon University [2]University of Pennsylvania [3]University of Washington
[4]M-A-P [5]The Hong Kong Polytechnic University

ziyuh@seas.upenn.edu yuetaili@uw.edu xyue2@andrew.cmu.edu

Abstract: Math reasoning has become the poster child of progress in large language models (LLMs), with new models rapidly surpassing human-level performance on benchmarks like MATH and AIME. But as math leaderboards improve week by week, it is worth asking: *do these gains reflect broader problem-solving ability or just narrow overfitting?* To answer this question, we evaluate over 20 open-weight reasoning-tuned models across a broad suite of tasks, including math, scientific QA, agent planning, coding, and standard instruction-following. We surprisingly find that most models that succeed in math fail to transfer their gains to other domains. To rigorously study this phenomenon, we conduct controlled experiments on Qwen3-14B models using math-only data but different tuning methods. We find that reinforcement learning (RL)-tuned models generalize well across domains, while supervised fine-tuning (SFT)-tuned models often forget general capabilities. Latent-space representation and token-space distribution shift analyses reveal that SFT induces substantial representation and output drift, while RL preserves general-domain structure. Our results suggest a need to rethink standard post-training recipes, particularly the reliance on SFT-distilled data for advancing reasoning models.

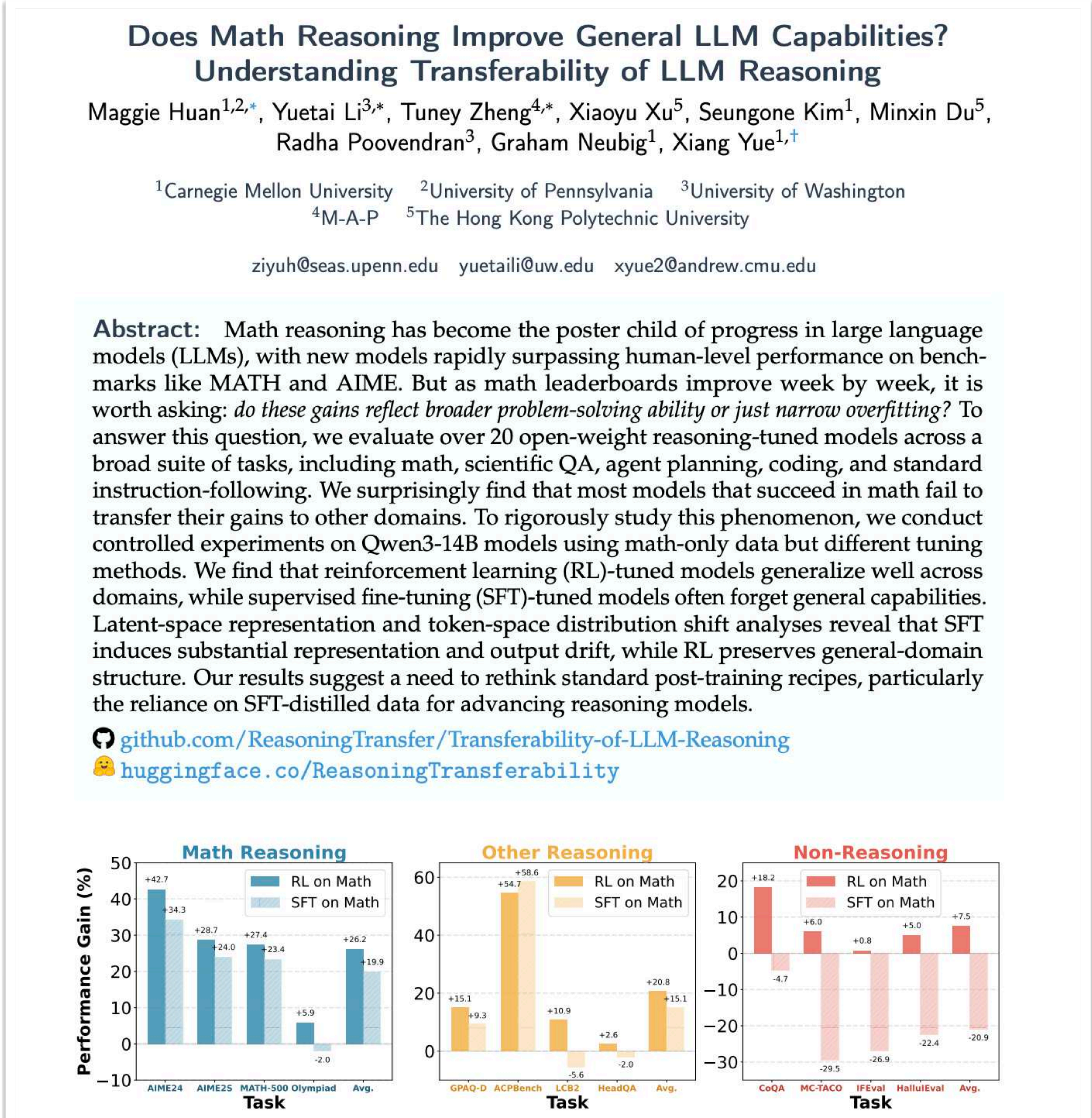
github.com/ReasoningTransfer/Transferability-of-LLM-Reasoning
🤗 huggingface.co/ReasoningTransferability



SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training

Tianzhe Chu[♠,*] Yuexiang Zhai[♥,♣,*] Jihan Yang[♦] Shengbang Tong[♦]
Saining Xie[♣,♦] Dale Schuurmans[♣,♛] Quoc V. Le[♣] Sergey Levine[♥] Yi Ma[♠,♥]

2025: The Rise of LRMs (as opposed to LLMs)



Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

Maggie Huan[1,2,*], Yuetai Li[3,*], Tuney Zheng[4,*], Xiaoyu Xu[5], Seungone Kim[1], Minxin Du[5], Radha Poovendran[3], Graham Neubig[1], Xiang Yue[1,†]

[1]Carnegie Mellon University [2]University of Pennsylvania [3]University of Washington
[4]M-A-P [5]The Hong Kong Polytechnic University

ziyuh@seas.upenn.edu yuetaili@uw.edu xyue2@andrew.cmu.edu

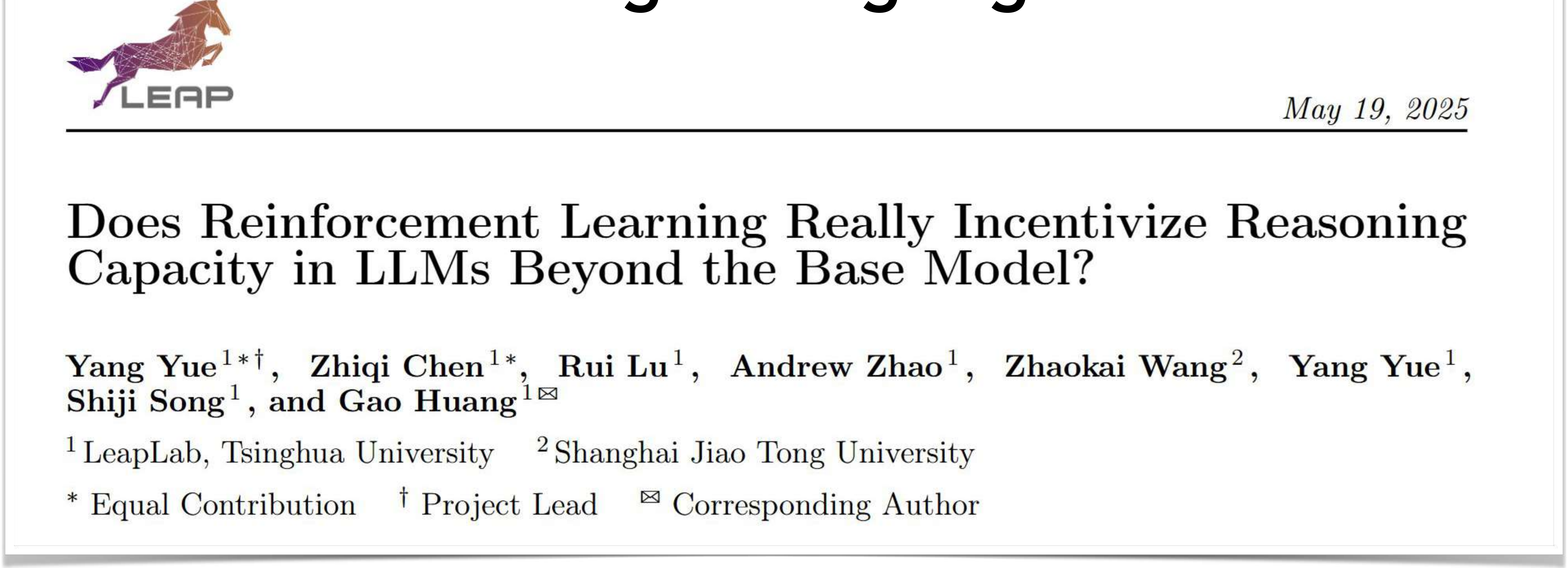
Abstract: Math reasoning has become the poster child of progress in large language models (LLMs), with new models rapidly surpassing human-level performance on benchmarks like MATH and AIME. But as math leaderboards improve week by week, it is worth asking: *do these gains reflect broader problem-solving ability or just narrow overfitting?* To answer this question, we evaluate over 20 open-weight reasoning-tuned models across a broad suite of tasks, including math, scientific QA, agent planning, coding, and standard instruction-following. We surprisingly find that most models that succeed in math fail to transfer their gains to other domains. To rigorously study this phenomenon, we conduct controlled experiments on Qwen3-14B models using math-only data but different tuning methods. We find that reinforcement learning (RL)-tuned models generalize well across domains, while supervised fine-tuning (SFT)-tuned models often forget general capabilities. Latent-space representation and token-space distribution shift analyses reveal that SFT induces substantial representation and output drift, while RL preserves general-domain structure. Our results suggest a need to rethink standard post-training recipes, particularly the reliance on SFT-distilled data for advancing reasoning models.

github.com/ReasoningTransfer/Transferability-of-LLM-Reasoning
huggingface.co/ReasoningTransferability

SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training

Tianzhe Chu♠* Yuexiang Zhai♥♠* Jihan Yang♦ Shengbang Tong♦
Saining Xie♠♦ Dale Schuurmans♠♣ Quoc V. Le♠ Sergey Levine♥ Yi Ma♠♥

Striking findings against RL



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

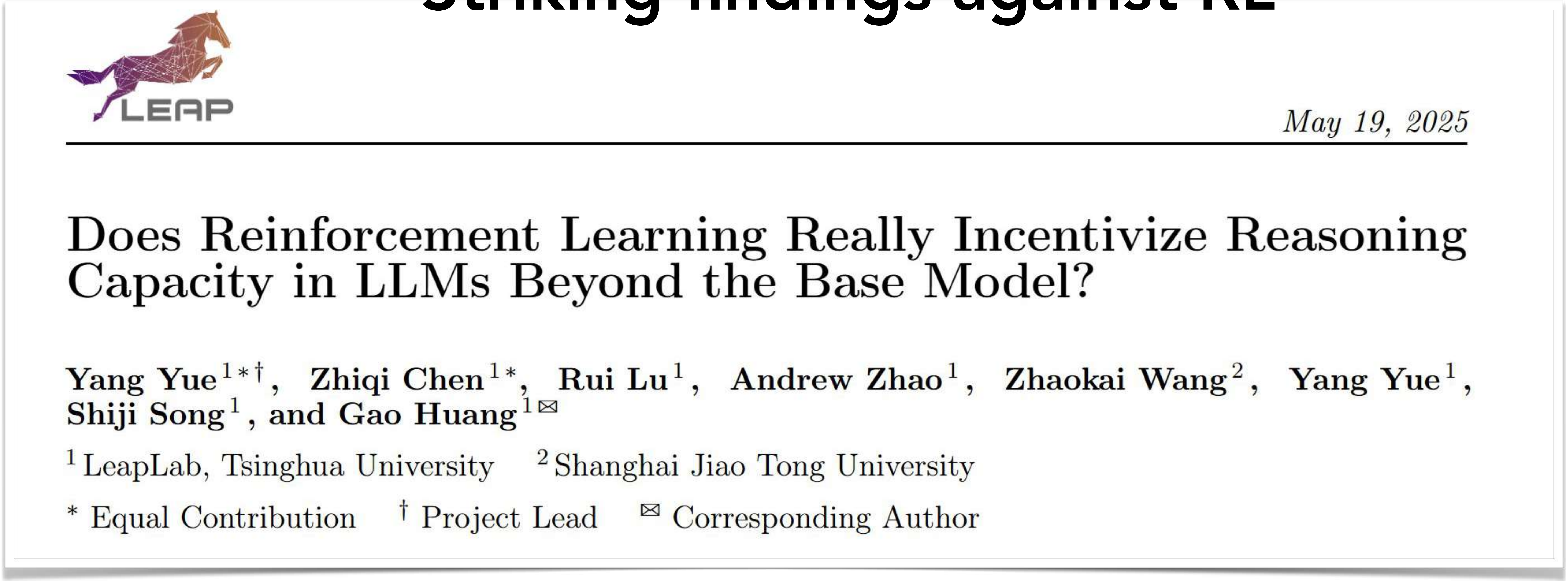
* Equal Contribution † Project Lead ✉ Corresponding Author

Pass@1 better after RLVR

Pass@K worse after RLVR

— compared to the base LLM —

Striking findings against RL



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

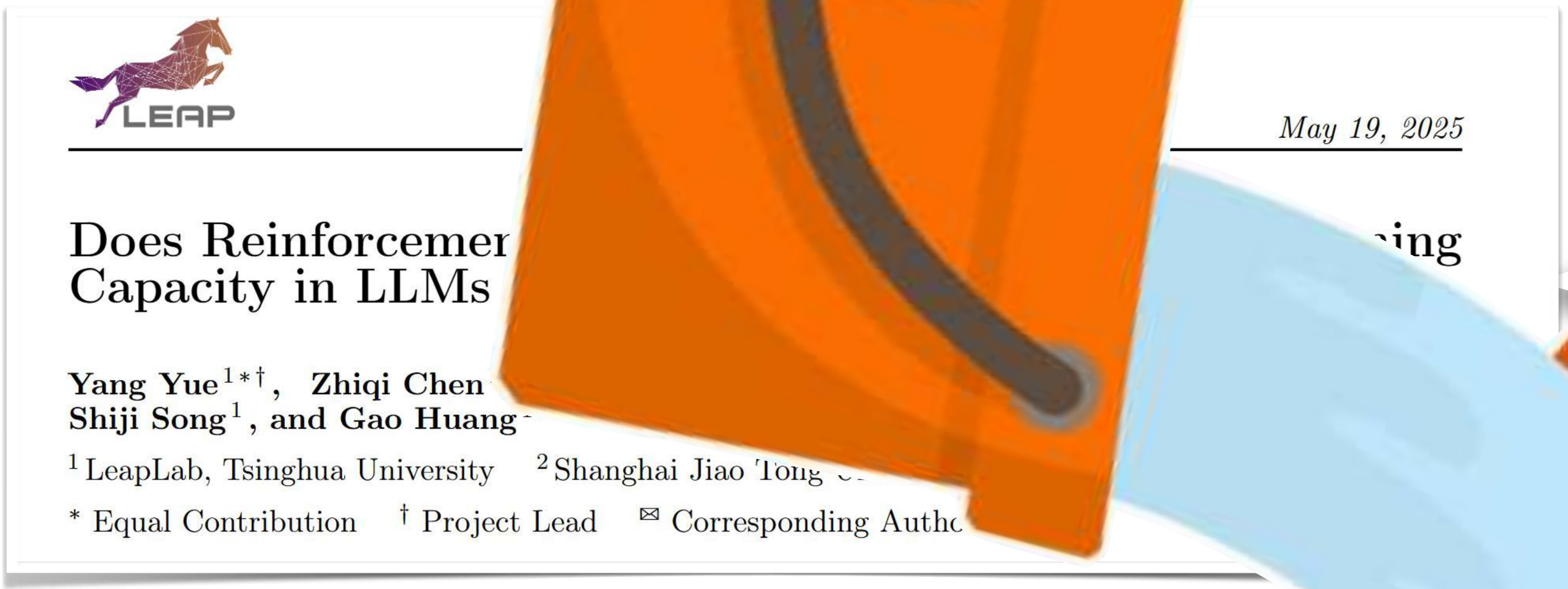
* Equal Contribution † Project Lead ✉ Corresponding Author

Pass@1 better after RLVR

Pass@K worse after RLVR

— compared to the base LLM —

St king fi ings against RL



LEAP

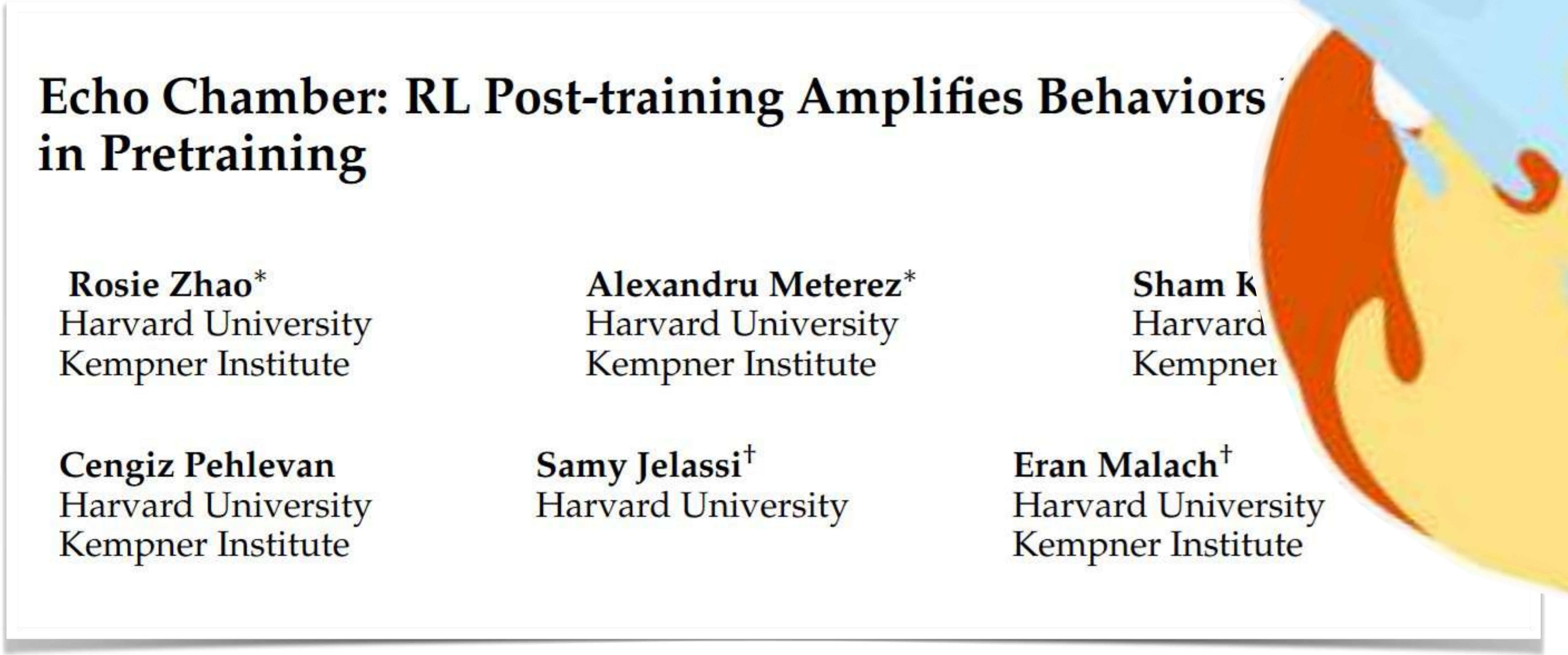
May 19, 2025

Does Reinforcemen ing
Capacity in LLMs

Yang Yue[1*†], Zhiqi Chen
Shiji Song[1], and Gao Huang

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

* Equal Contribution † Project Lead ✉ Corresponding Autho



Echo Chamber: RL Post-training Amplifies Behaviors
in Pretraining

Rosie Zhao*
Harvard University
Kempner Institute

Alexandru Meterez*
Harvard University
Kempner Institute

Sham K
Harvard
Kempner

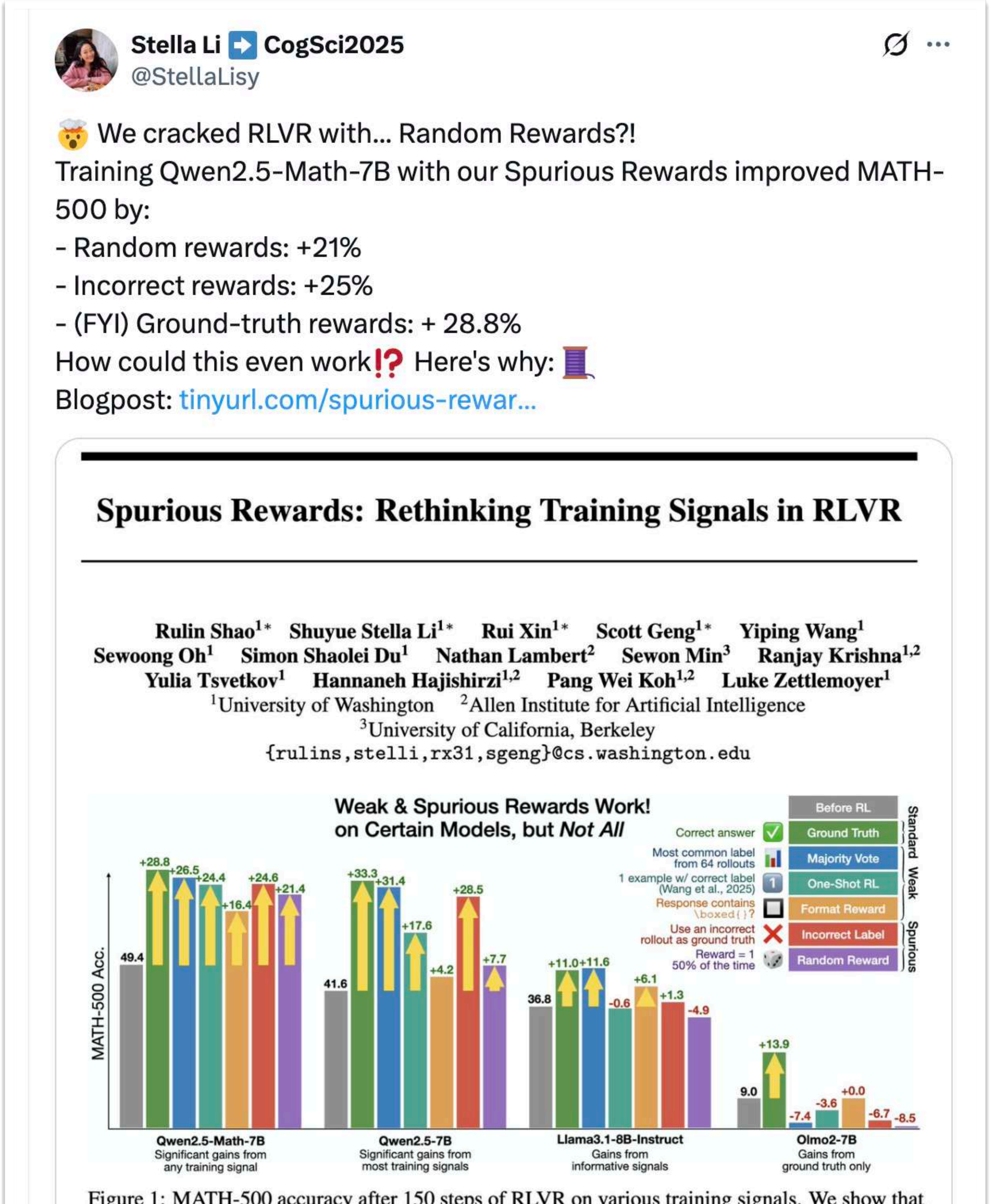
Cengiz Pehlevan
Harvard University
Kempner Institute

Samy Jelassi†
Harvard University

Eran Malach†
Harvard University
Kempner Institute

Pas better after RLVR
Pa worse after RLVR
— c ed to the base LLM —

Striking findings against RL



Stella Li ➡️ CogSci2025
@StellaLisy

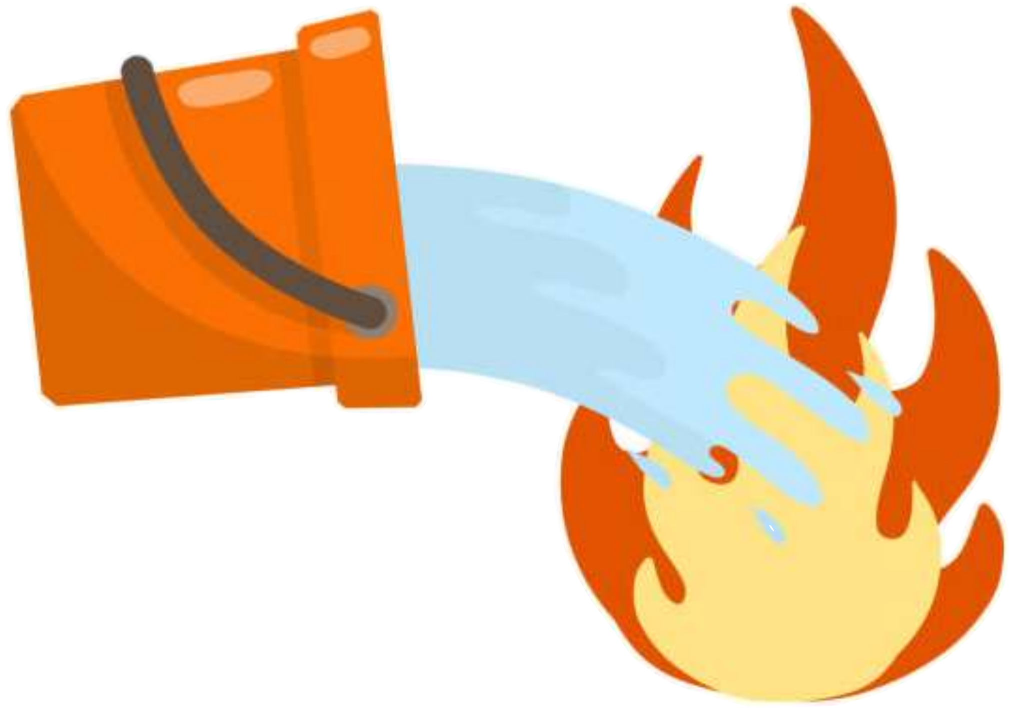
🤯 We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved MATH-500 by:
- Random rewards: +21%
- Incorrect rewards: +25%
- (FYI) Ground-truth rewards: + 28.8%
How could this even work⁉️ Here's why: 🧵
Blogpost: tinyurl.com/spurious-rewar...

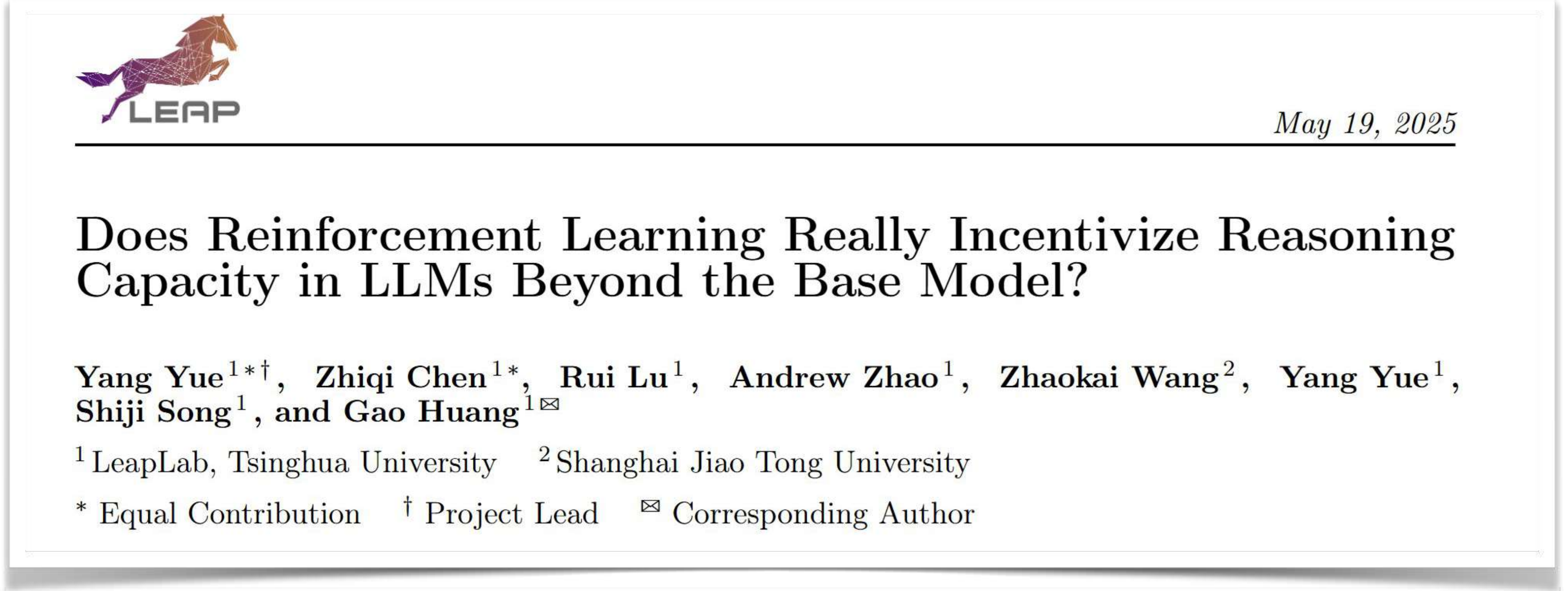
Spurious Rewards: Rethinking Training Signals in RLVR

Rulin Shao[1]* Shuyue Stella Li[1]* Rui Xin[1]* Scott Geng[1]* Yiping Wang[1]
Sewoong Oh[1] Simon Shaolei Du[1] Nathan Lambert[2] Sewon Min[3] Ranjay Krishna[1,2]
Yulia Tsvetkov[1] Hannaneh Hajishirzi[1,2] Pang Wei Koh[1,2] Luke Zettlemoyer[1]
[1]University of Washington [2]Allen Institute for Artificial Intelligence
[3]University of California, Berkeley
{rulins,stelli,rx31,sgeng}@cs.washington.edu

Weak & Spurious Rewards Work!
on Certain Models, but *Not All*
Before RL
Correct answer ✅ Ground Truth
Most common label from 64 rollouts 📊 Majority Vote
1 example w/ correct label (Wang et al., 2025) 1 One-Shot RL
Response contains \boxed{}? □ Format Reward
Use an incorrect rollout as ground truth ❌ Incorrect Label
Reward = 1 50% of the time 🎲 Random Reward
Standard
Weak
Spurious
MATH-500 Acc.
49.4
+28.8
+26.5
+24.4
+16.4
+24.6
+21.4
41.6
+33.3
+31.4
+17.6
+4.2
+28.5
+7.7
36.8
+11.0
+11.6
-0.6
+6.1
+1.3
-4.9
9.0
+13.9
-7.4
-3.6
+0.0
-6.7
-8.5
Qwen2.5-Math-7B
Significant gains from any training signal
Qwen2.5-7B
Significant gains from most training signals
Llama3.1-8B-Instruct
Gains from informative signals
Olmo2-7B
Gains from ground truth only

Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that



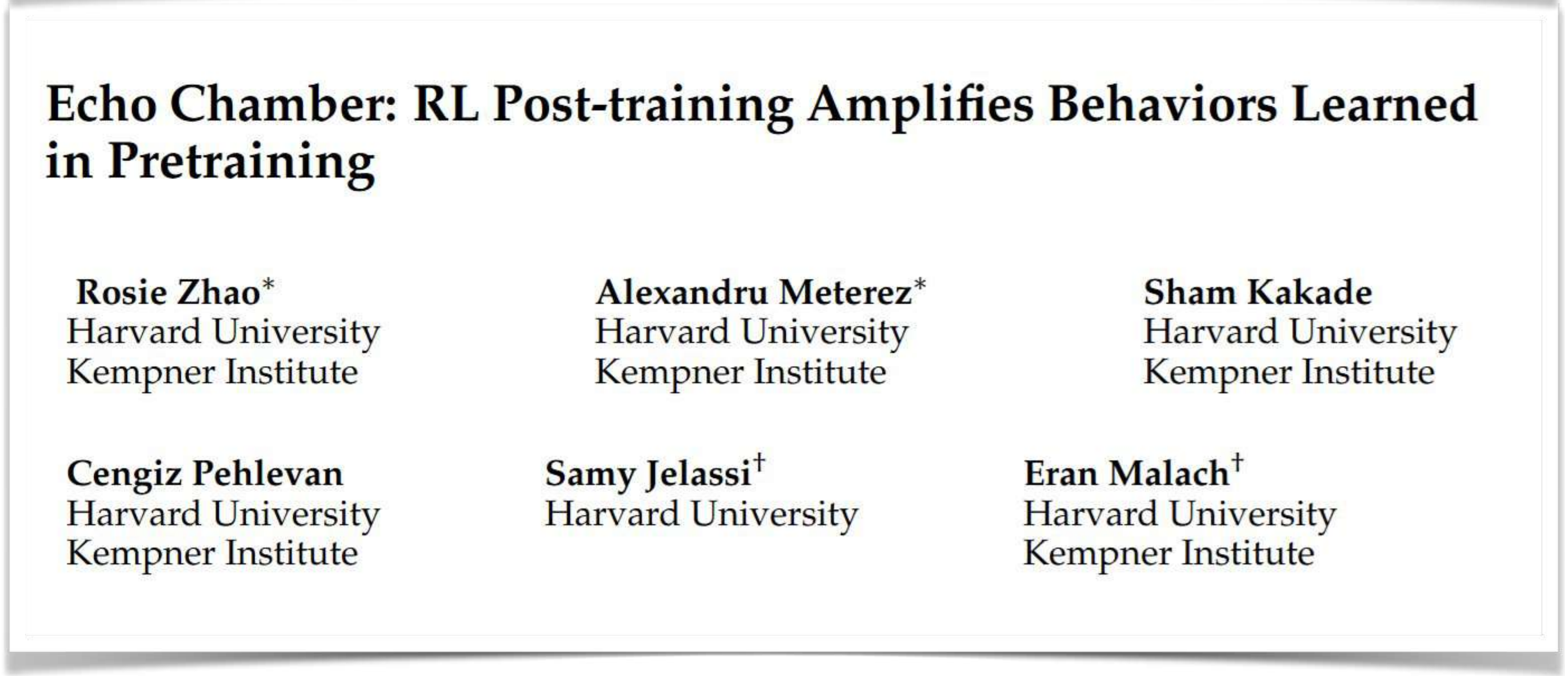


LEAP

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



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

Rosie Zhao*
Harvard University
Kempner Institute

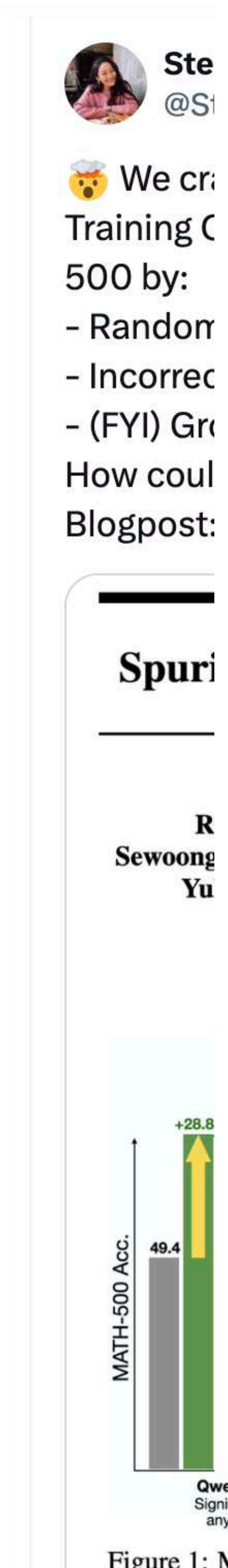
Alexandru Meterez*
Harvard University
Kempner Institute

Sham Kakade
Harvard University
Kempner Institute

Cengiz Pehlevan
Harvard University
Kempner Institute

Samy Jelassi†
Harvard University

Eran Malach†
Harvard University
Kempner Institute



Ste
@S

😮 We cr
Training C
500 by:
- Random
- Incorrec
- (FYI) Gr
How coul
Blogpost:

Spuri

R
Sewoong
Yu

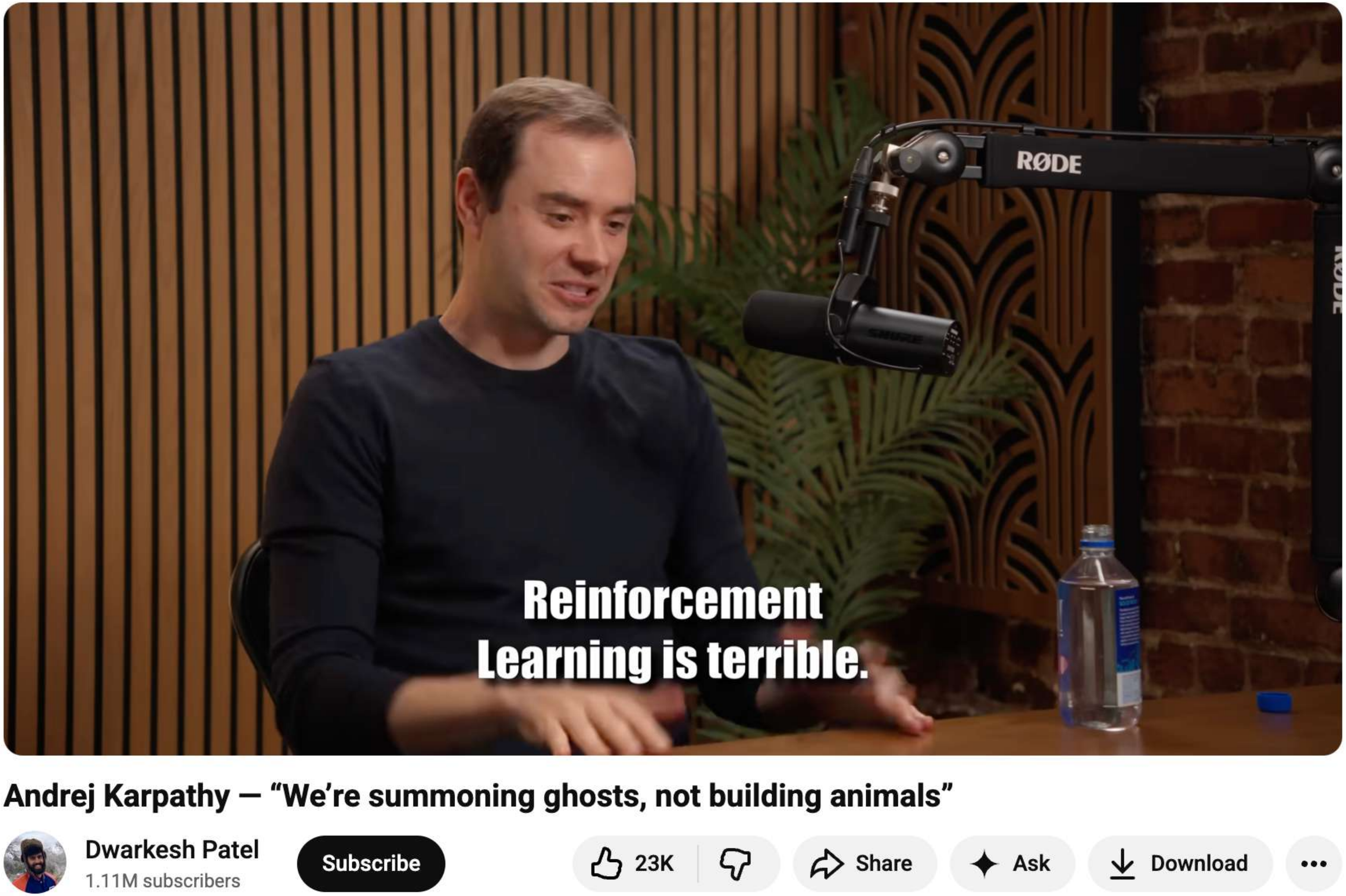
Qwen2.5-Math-7B
Significant gains from any training signal

Qwen2.5-7B
Significant gains from most training signals

Llama3.1-8B-Instruct
Gains from informative signals

Olmo2-7B
Gains from ground truth only

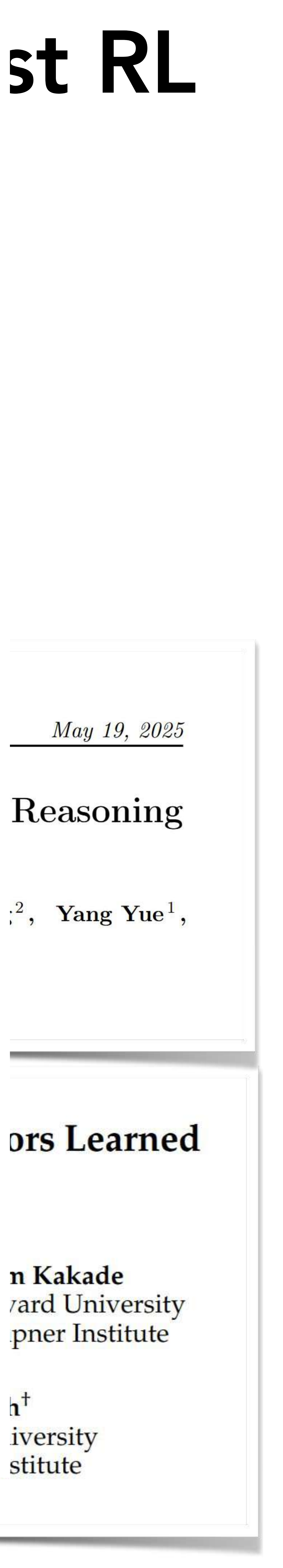
Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that



Reinforcement Learning is terrible.

Andrej Karpathy — “We’re summoning ghosts, not building animals”

Dwarkesh Patel
1.11M subscribers



st RL

May 19, 2025

Reasoning

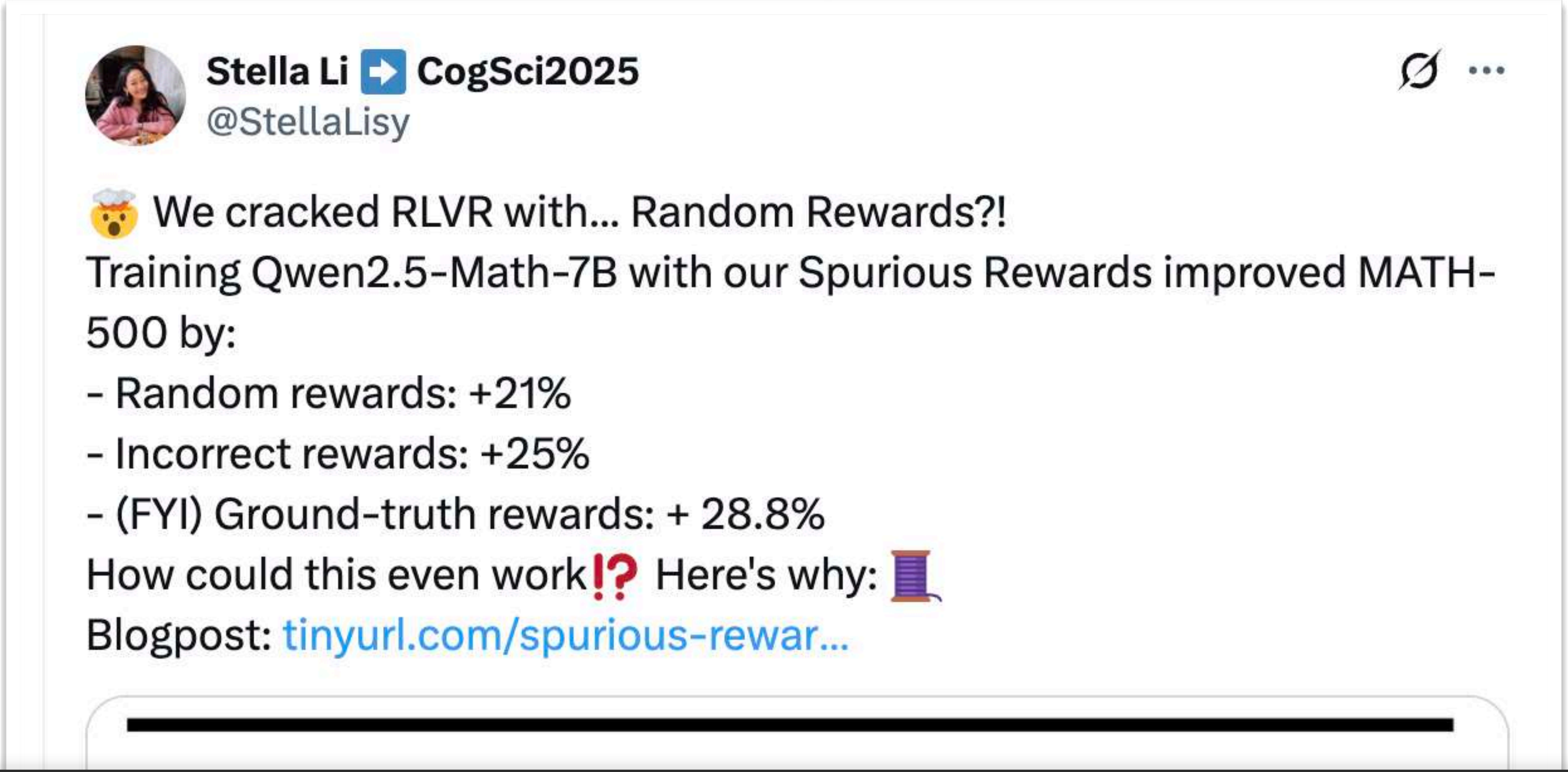
2, Yang Yue1,

ors Learned

n Kakade
ard University
pner Institute

h†
iversity
stitute

Striking findings against RL



Stella Li ➡️ CogSci2025
@StellaLisy

🤯 We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved MATH-500 by:
- Random rewards: +21%
- Incorrect rewards: +25%
- (FYI) Ground-truth rewards: + 28.8%
How could this even work⁉️ Here's why: 🧵
Blogpost: tinyurl.com/spurious-rewar...

Chemistry between the base LLM and RL matters

Conclusions from effortless RL != effortful RL

{rulins,stelli,rx31,sgeng}@cs.washington.edu

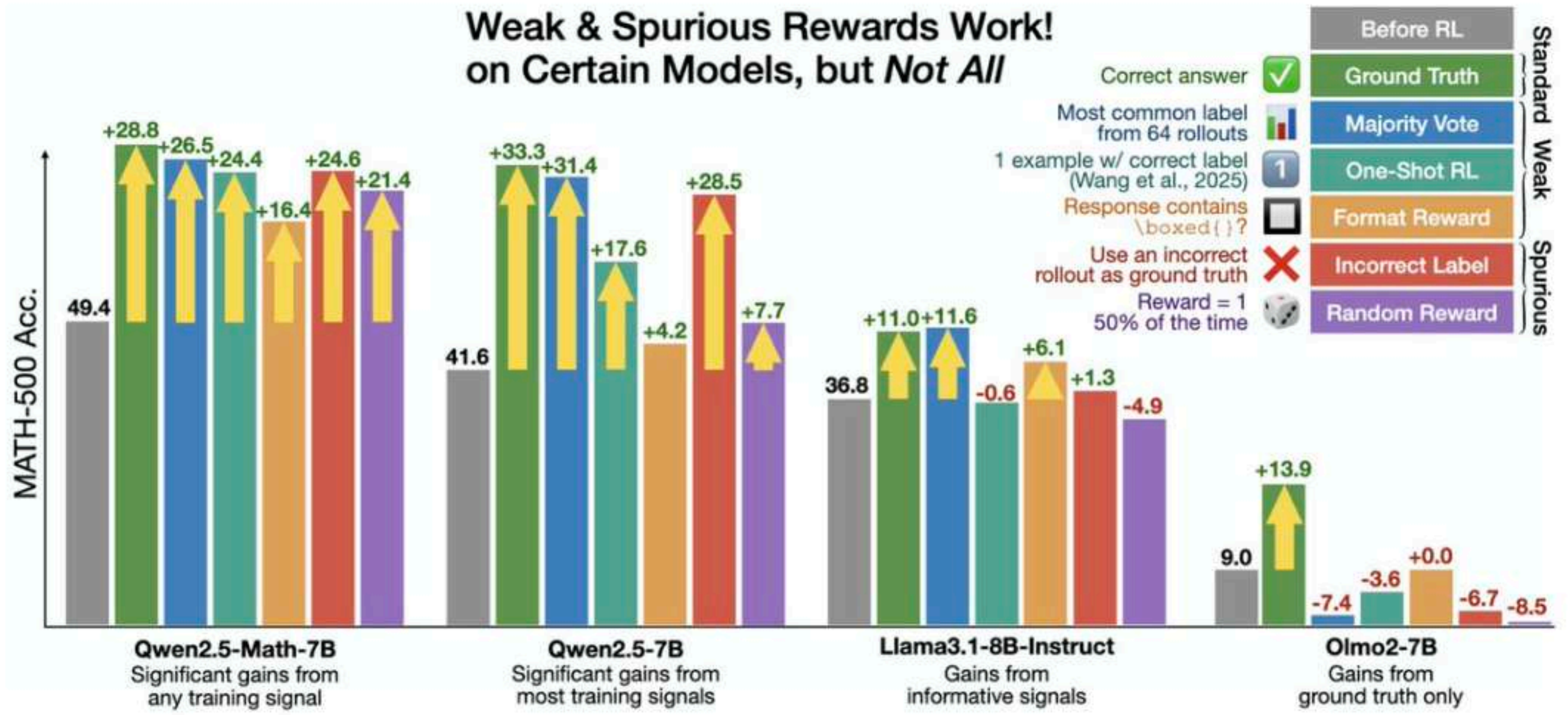
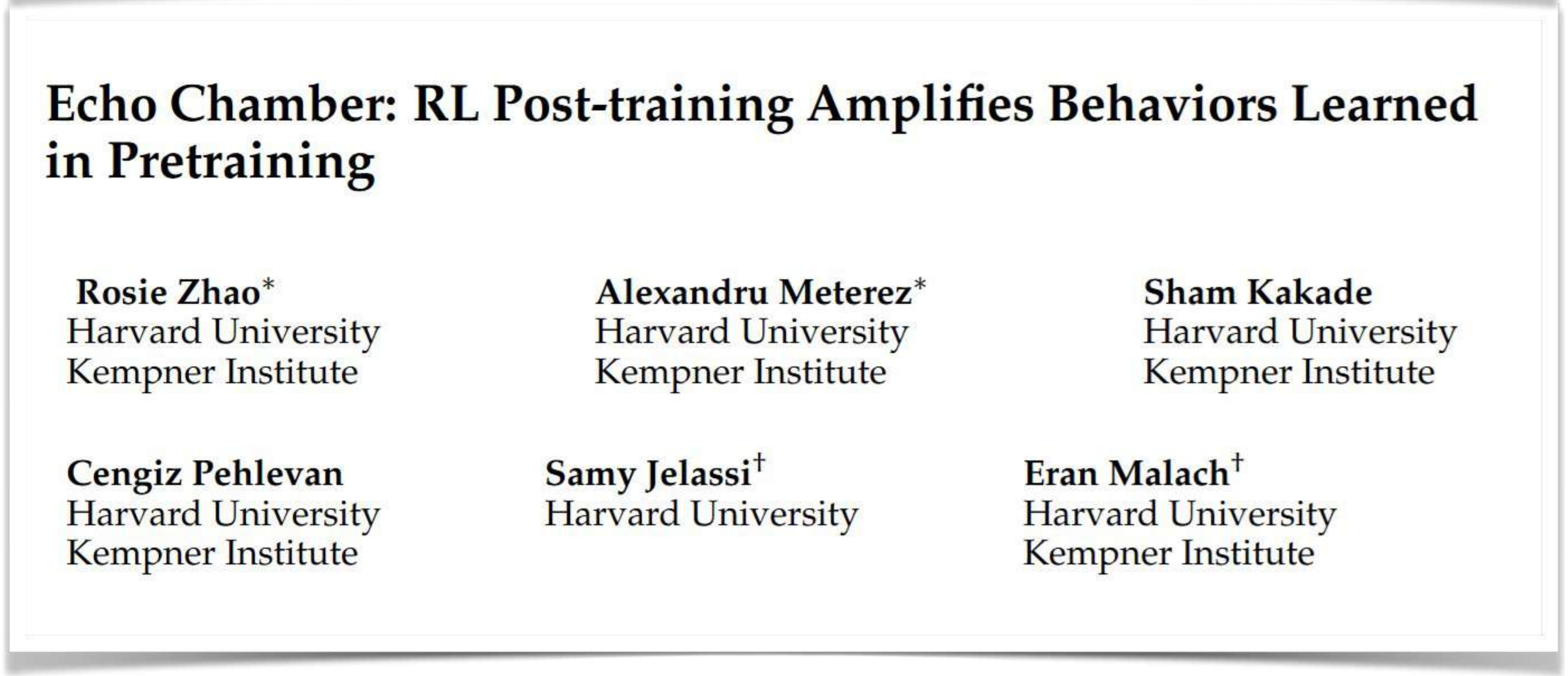


Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that

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



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

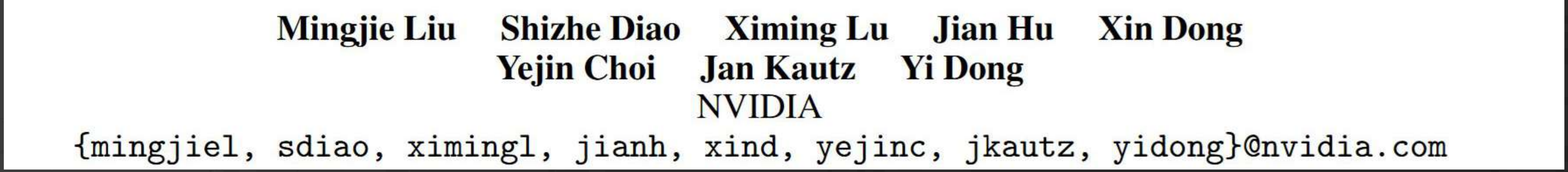
ProRL

“Rome wasn't built in a day”

What if we go for endurance with a small model (1.5B) rather than a sprint with a large model

— NeurIPS 2025 —

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

ByteDance

Prolonged Reinforcement Learning

Deepseek

ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

ByteDance

Prolonged Reinforcement Learning

Deepseek

ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

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

- **Dynamic Sampling** - targets mid-difficulty examples to maintain diverse learning signals
- **Decoupled Clipping** - 'clip-higher' uplifts low-probability tokens, encourages broader exploration.

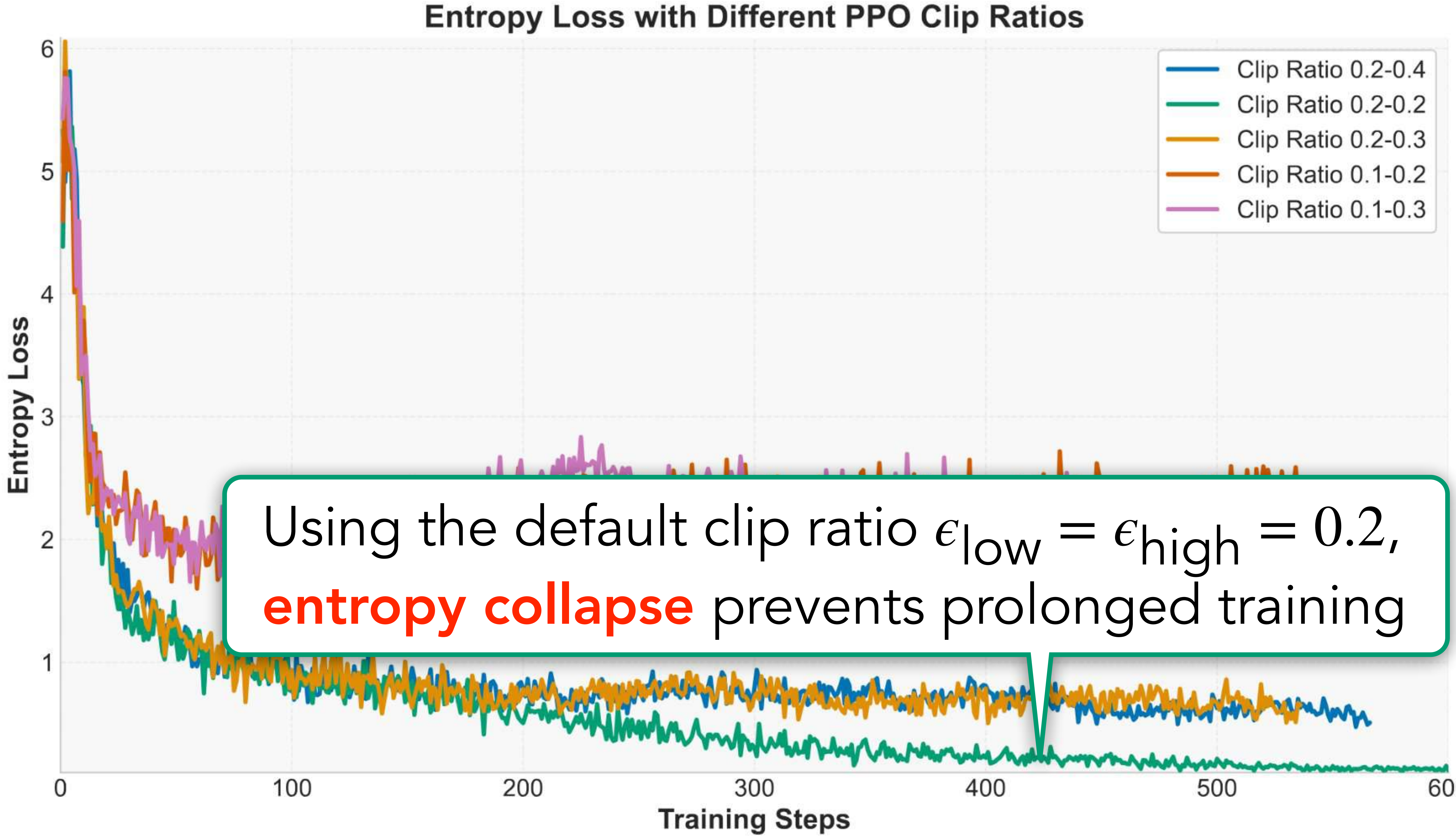
$$\text{clip}(r_\theta(\tau),\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}})$$

Prolonged Reinforcement Learning

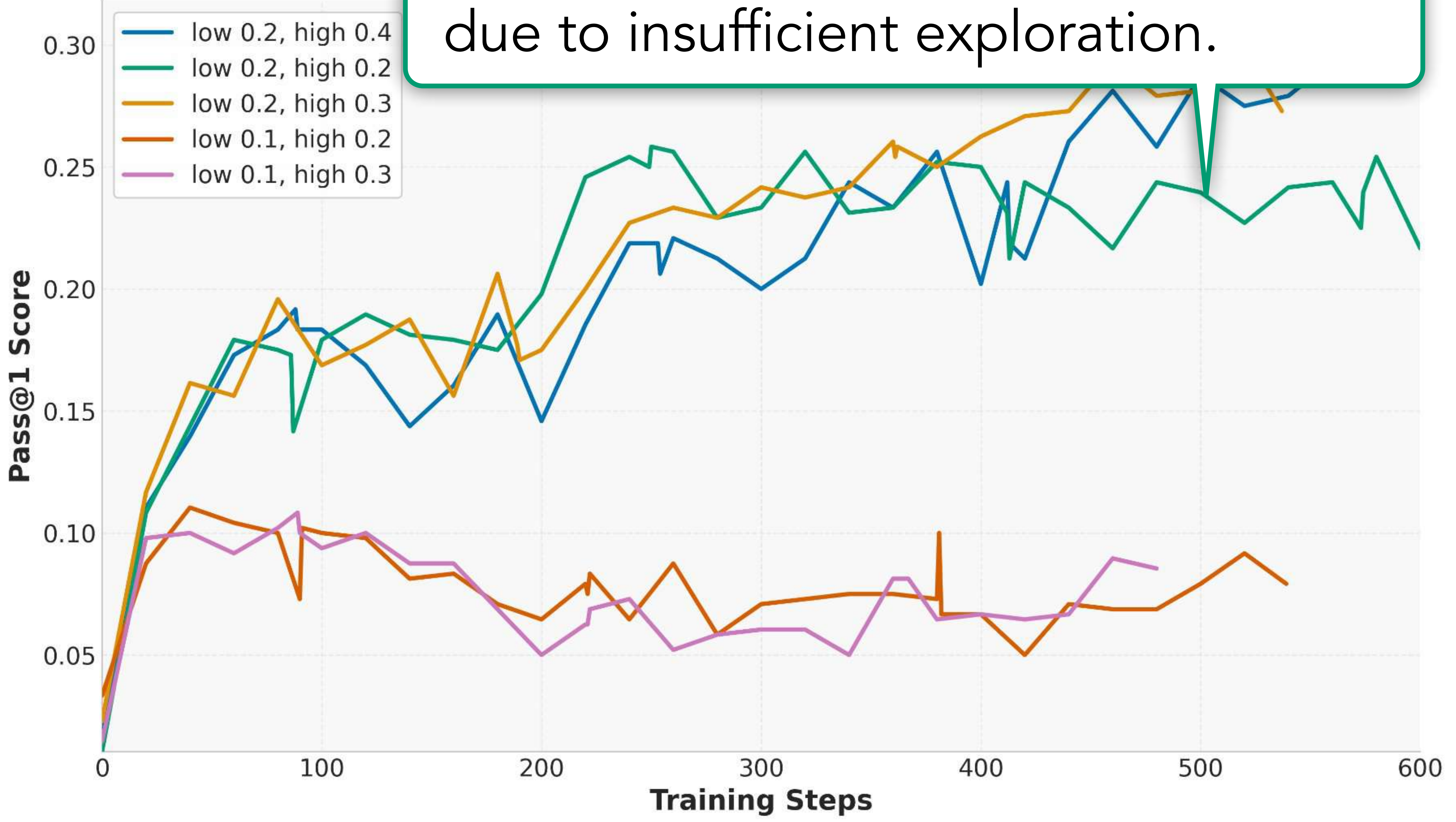
ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

Lesson I: Maintain **sustainable entropy** via clip ratio, balance exploration and exploitation.

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



Using the default clip ratio $\epsilon_{\text{low}} = \epsilon_{\text{high}} = 0.2$, **entropy collapse** prevents prolonged training



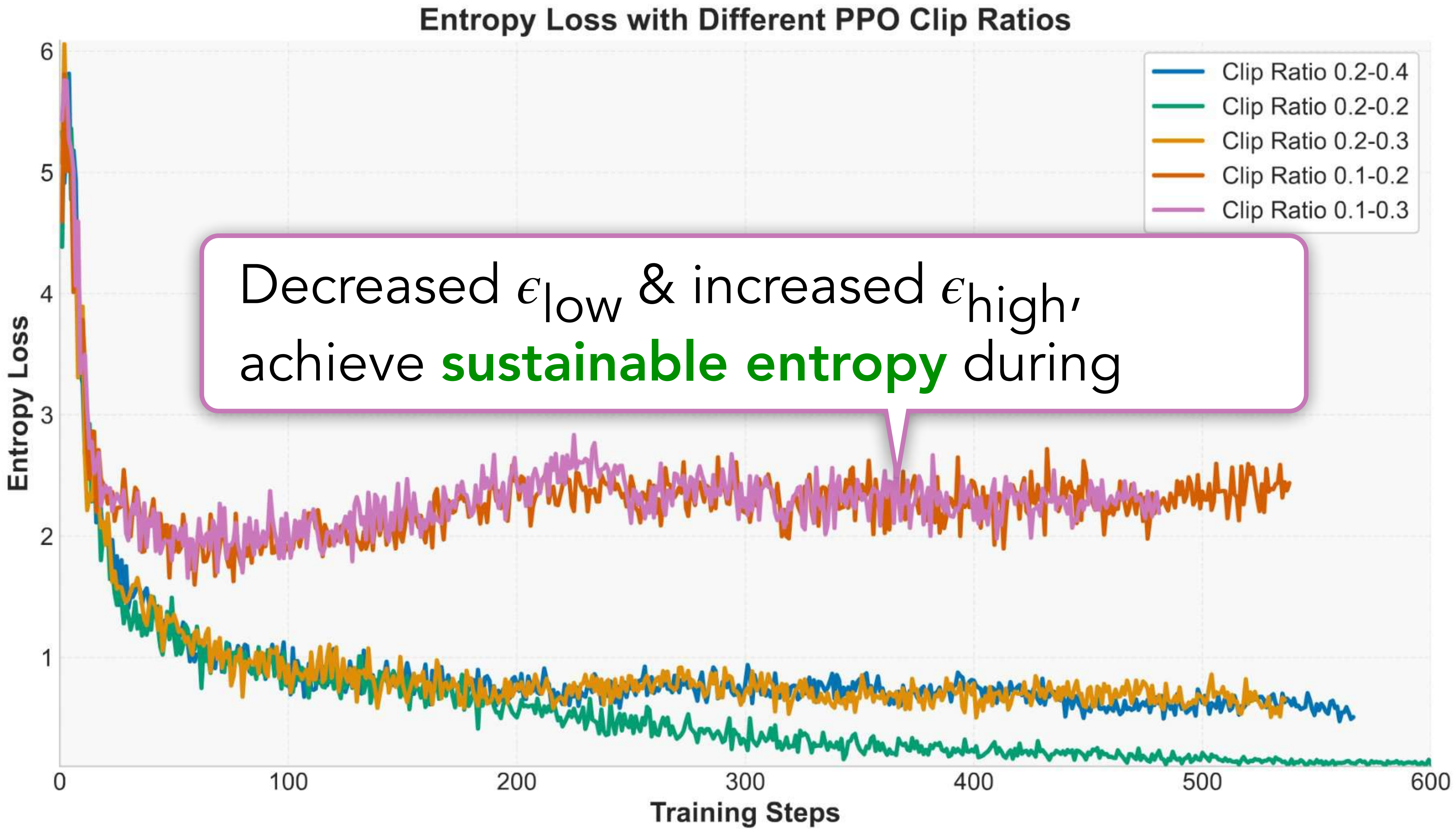
Performance initially increases but plateaus after entropy collapse, likely due to insufficient exploration.

Prolonged Reinforcement Learning

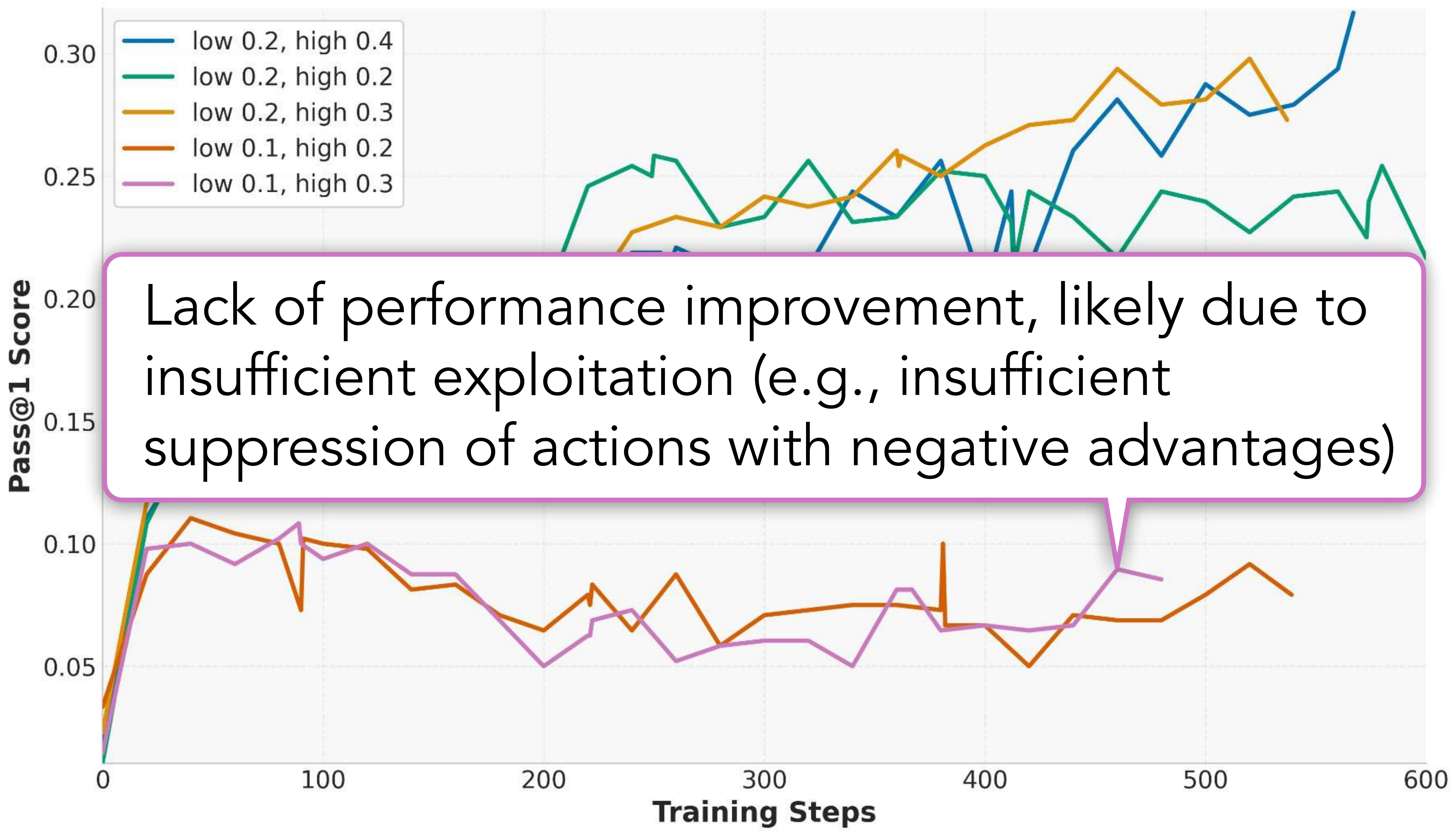
ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

Lesson I: Maintain **sustainable entropy** via clip ratio, balance exploration and exploitation.

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



Decreased ϵ_{low} & increased ϵ_{high}, achieve **sustainable entropy** during



Lack of performance improvement, likely due to insufficient exploitation (e.g., insufficient suppression of actions with negative advantages)

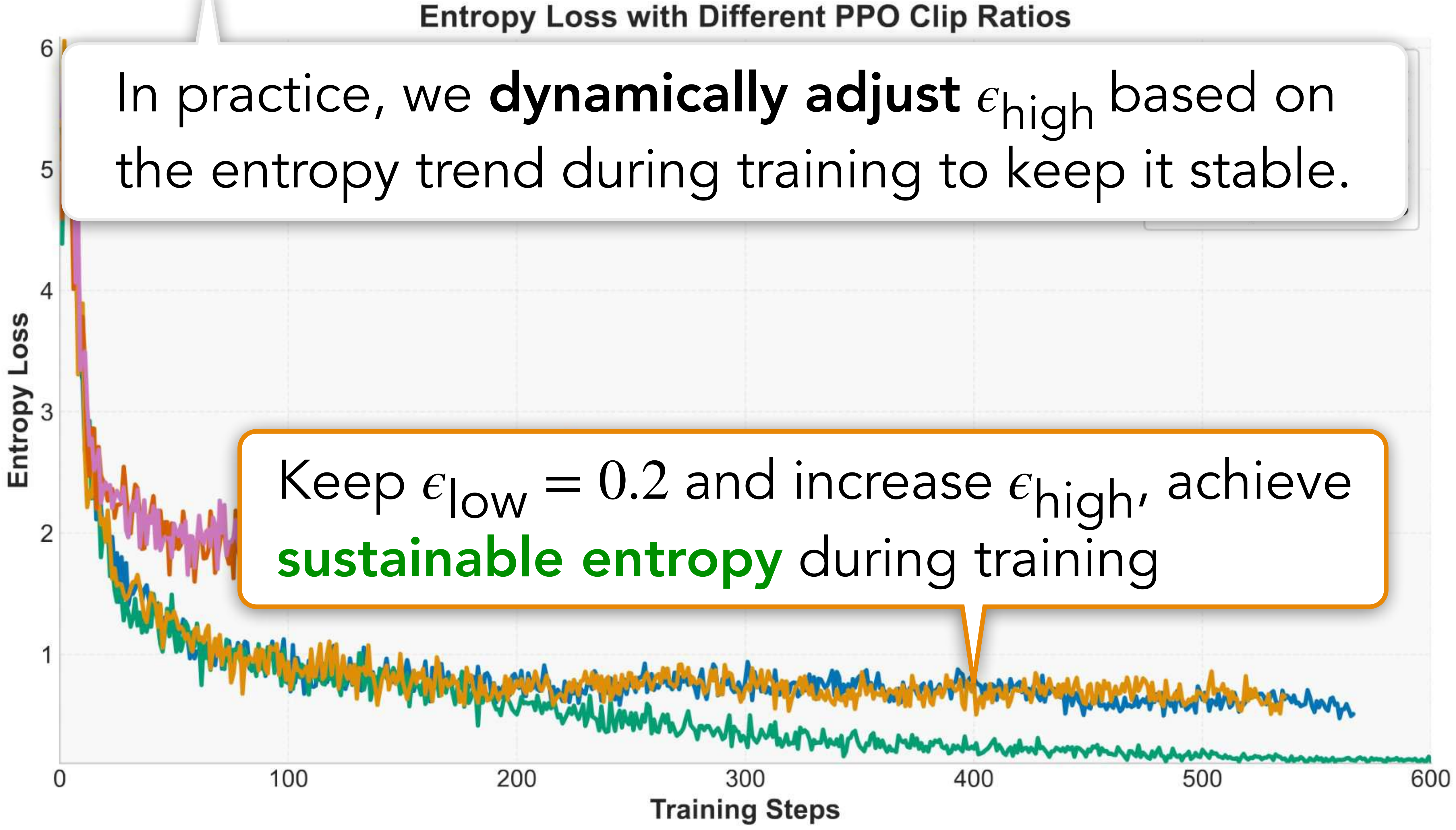
Prolonged Reinforcement Learning

ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

Lesson I: Maintain **sustainable entropy** via clip ratio, balance exploration and exploitation.

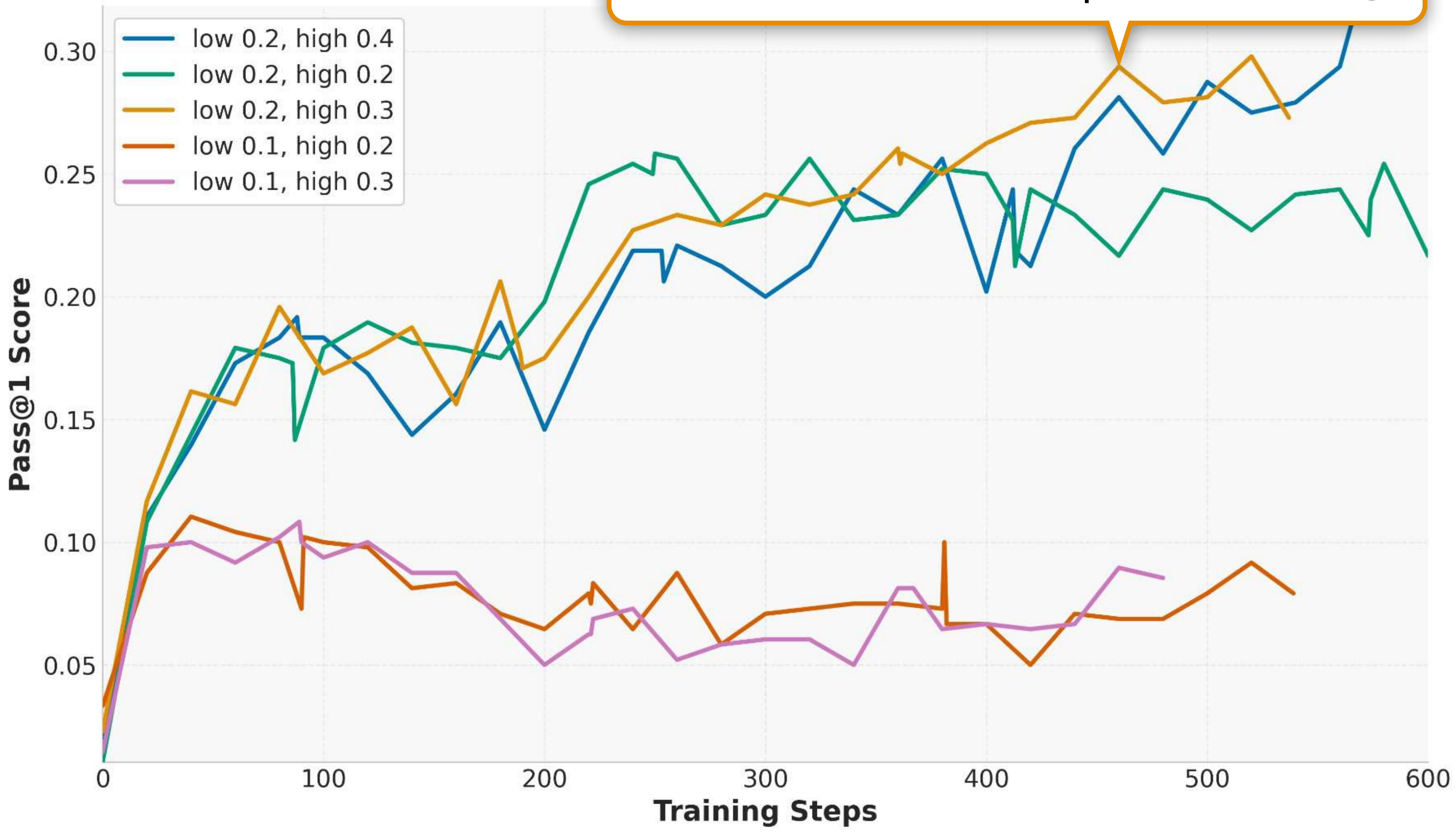
$$\text{clip}(r_\theta(\tau),\ 1-\epsilon_{\text{low}},\ 1+\epsilon_{\text{high}})$$

In practice, we **dynamically adjust** ϵ_{high} based on the entropy trend during training to keep it stable.



Keep $\epsilon_{\text{low}} = 0.2$ and increase ϵ_{high}, achieve **sustainable entropy** during training

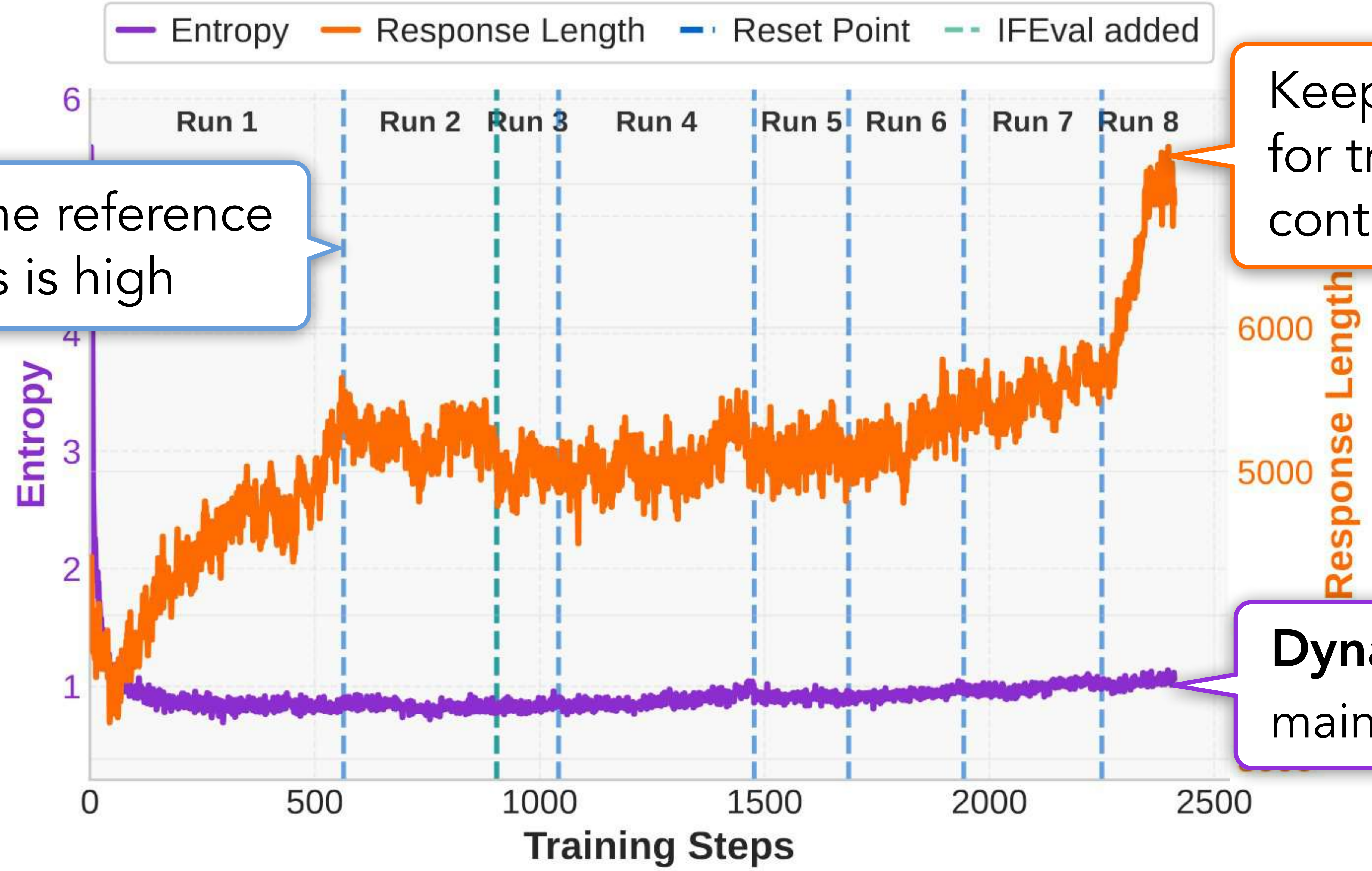
Performance keep increasing



Prolonged Reinforcement Learning

ProRL is built on top of one of the best practices in RL—Decoupled Clip and Dynamic Sampling Policy Optimization (**DAPO**), a variation of Group Relative Policy Optimization (GRPO).

Lesson III: Dynamically adjust hyper-parameters to **keep one hero run alive**



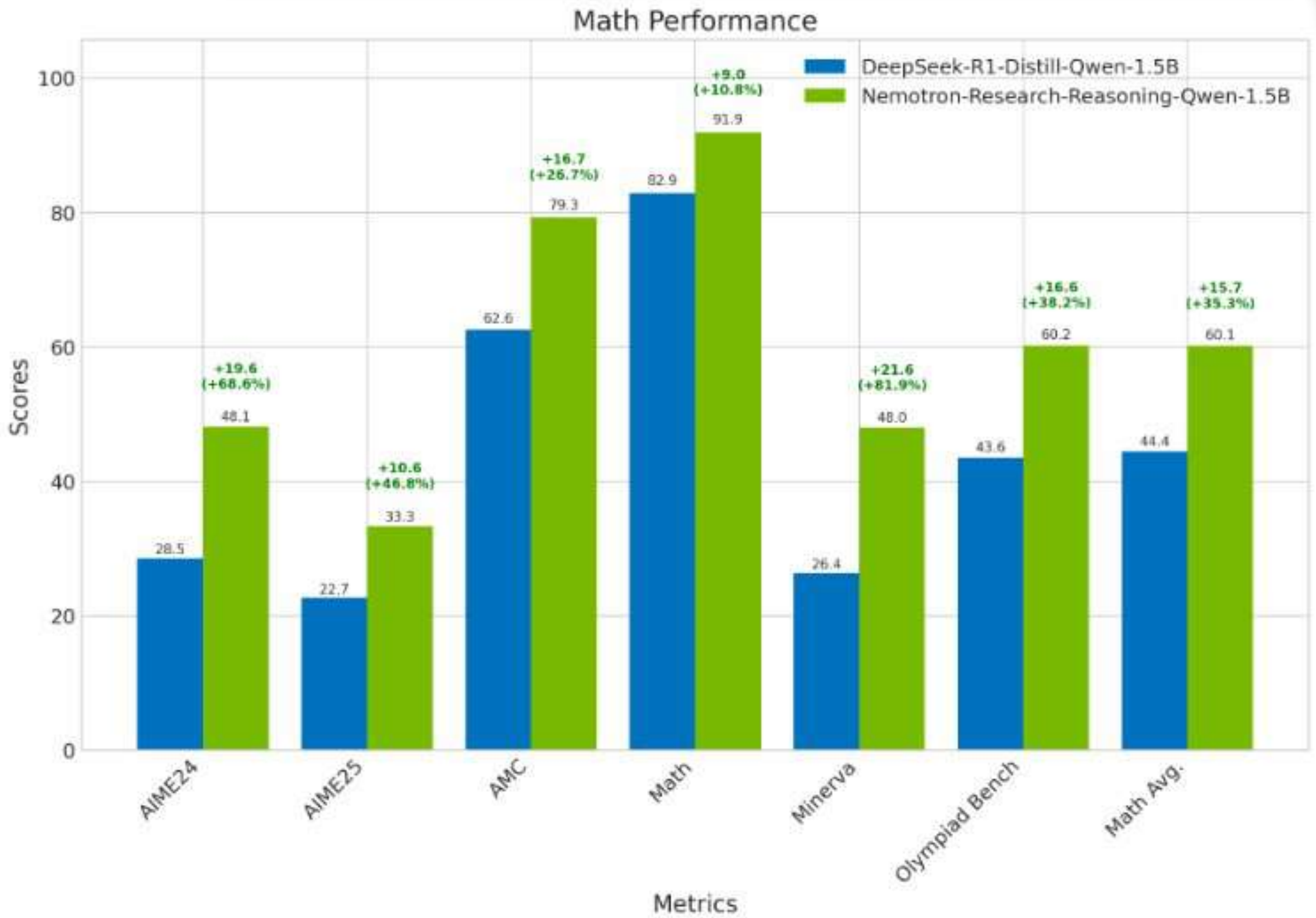
Periodically reset the reference policy when KL loss is high

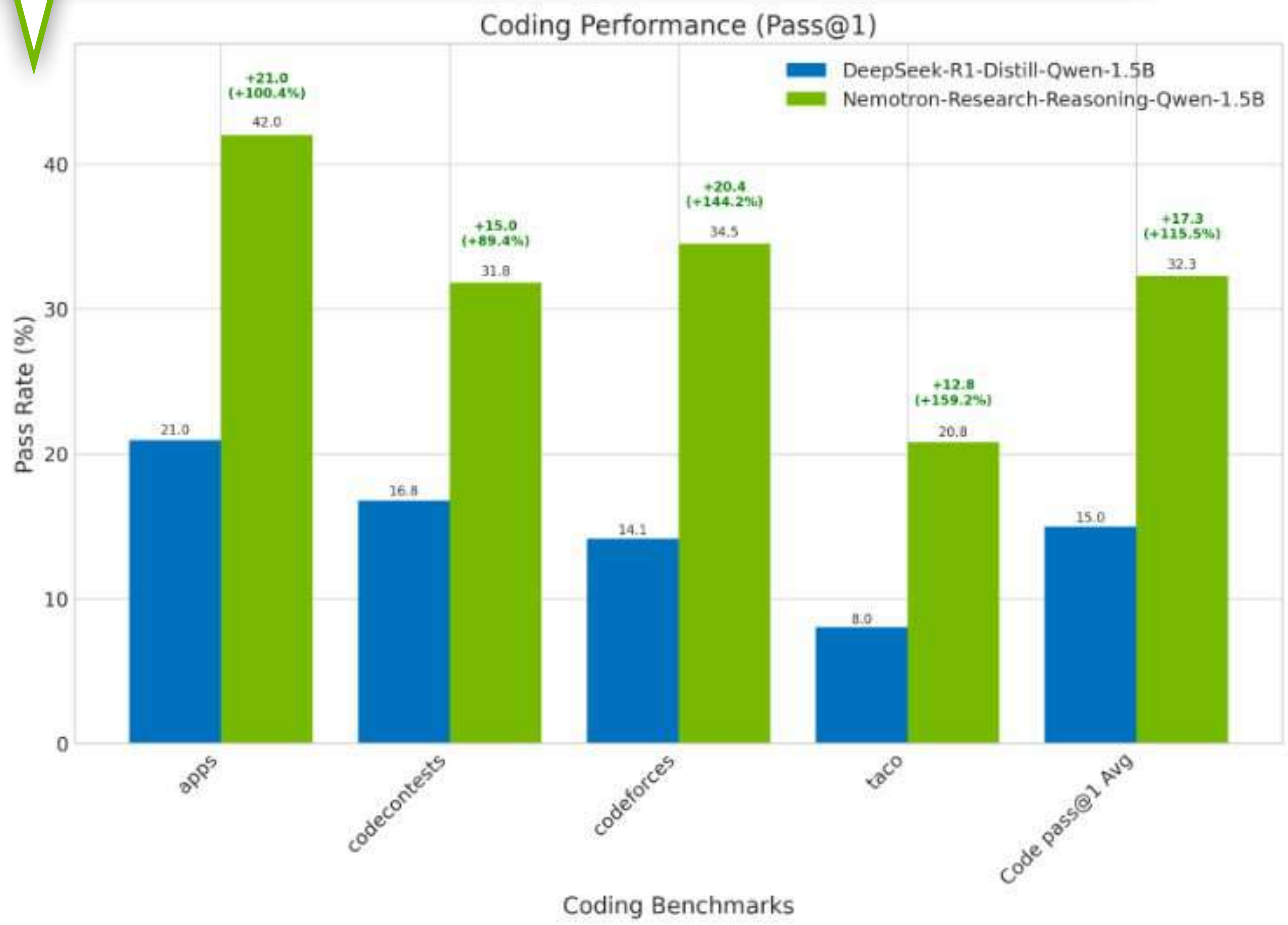
Keep response length under 8k for training efficiency; extend context length in the final run.

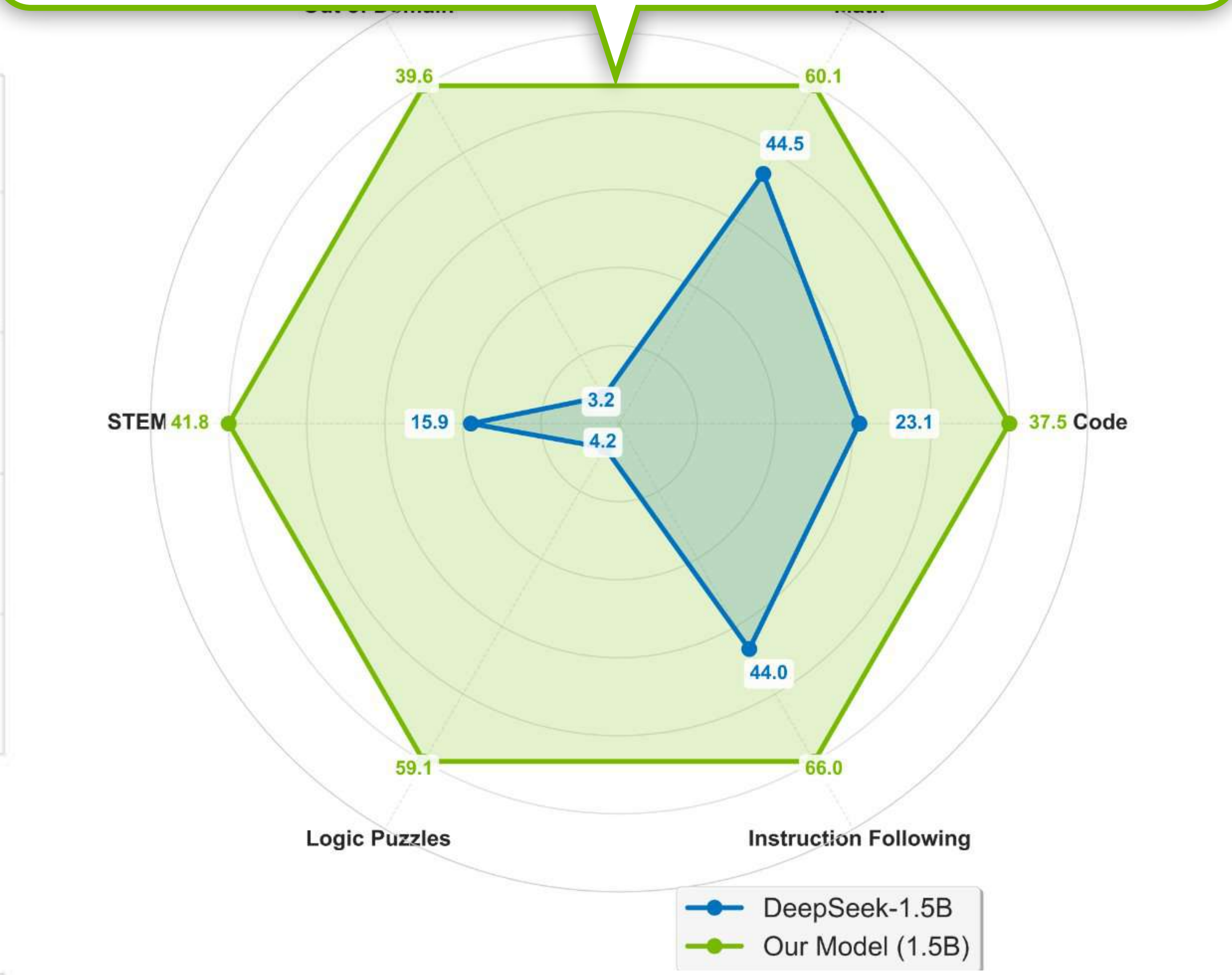
Dynamically adjust ϵ_{high} to maintain **sustainable entropy**

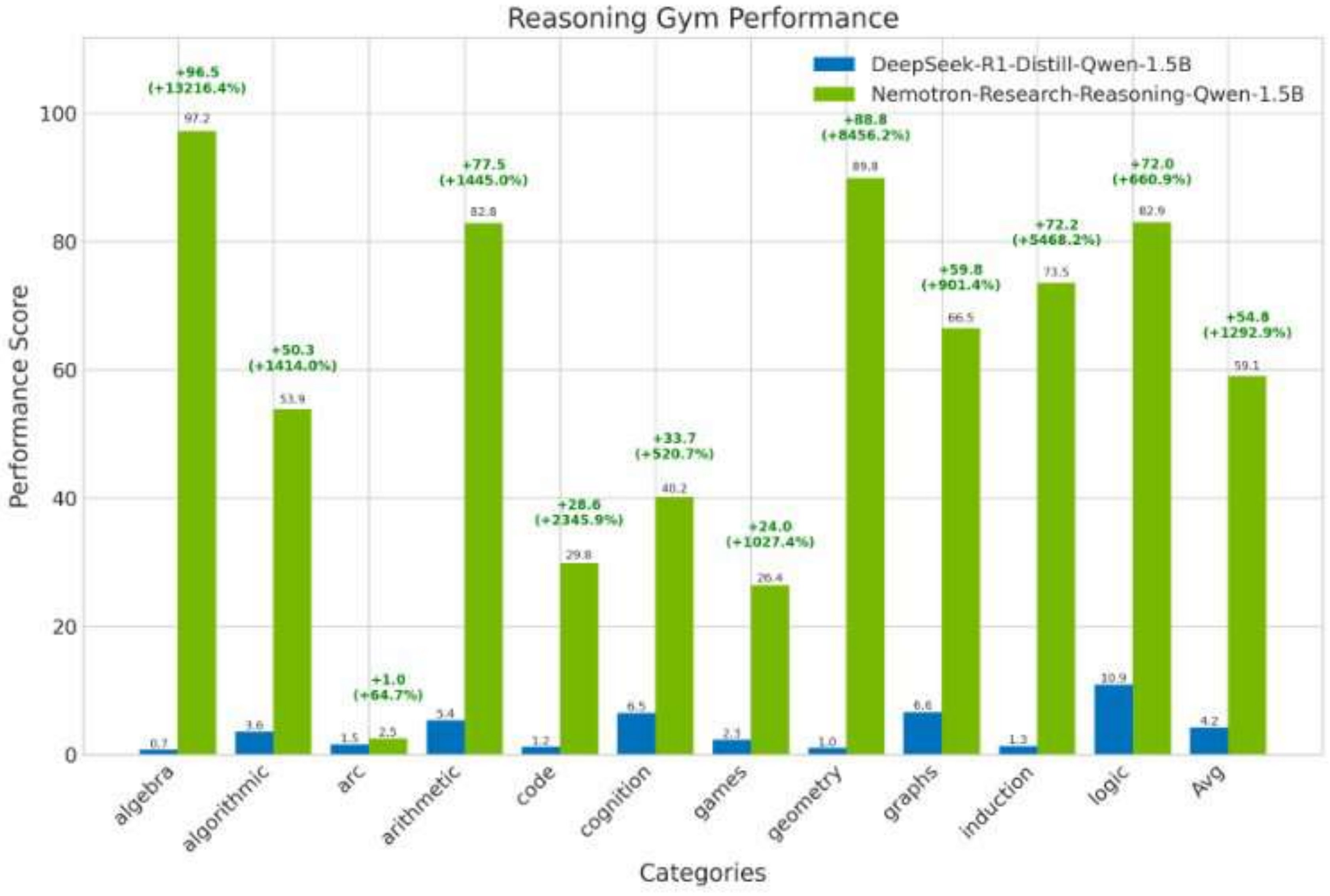
Evaluated across diverse domains—math, code, STEM, reasoning gym, and instruction following—ProRL **results in significant gains** over the base model, **Distill-R1-1.5B**, in all cases.

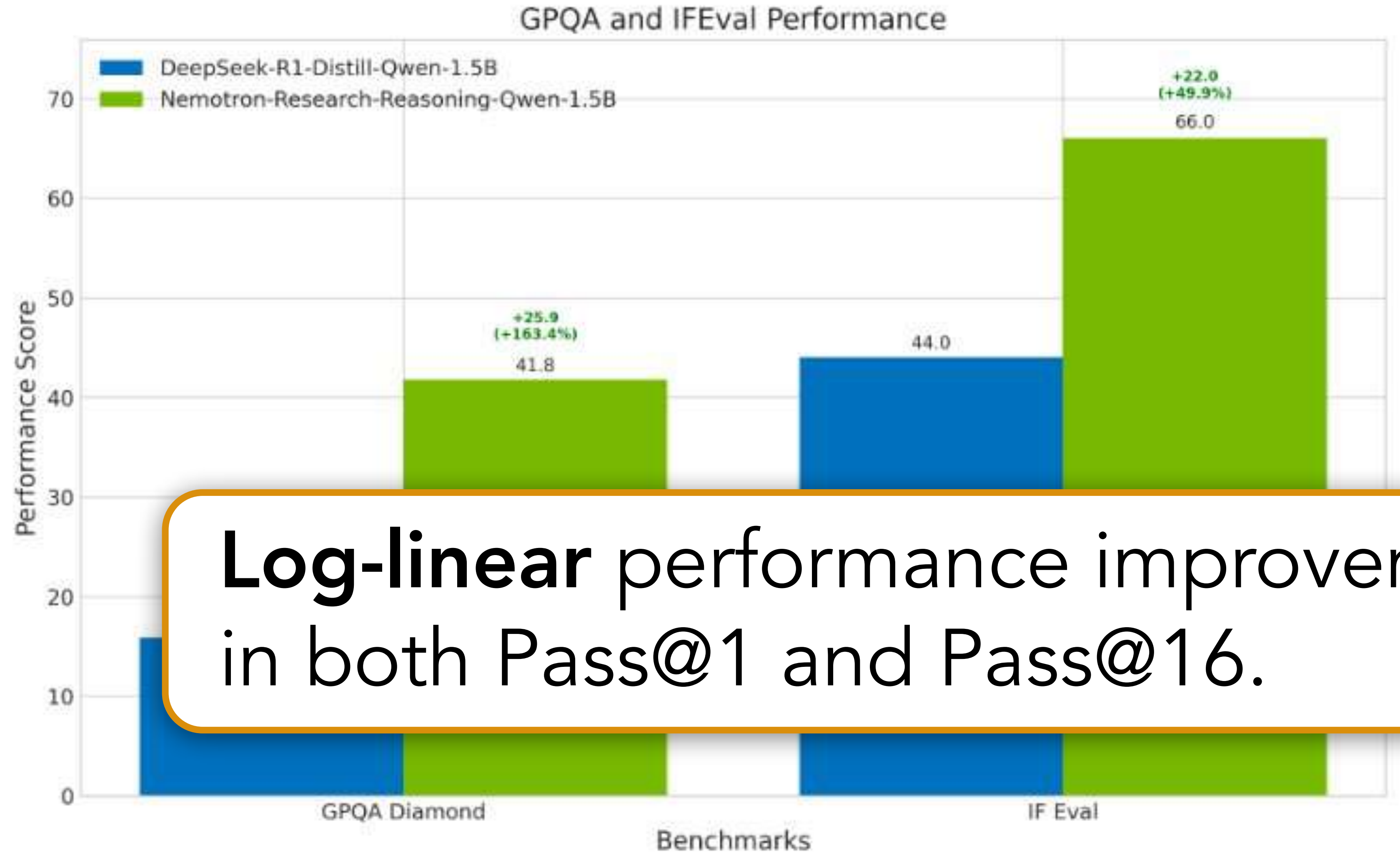
The gain is particularly large in out-of-domain tasks and reasoning gym where the base model struggles.

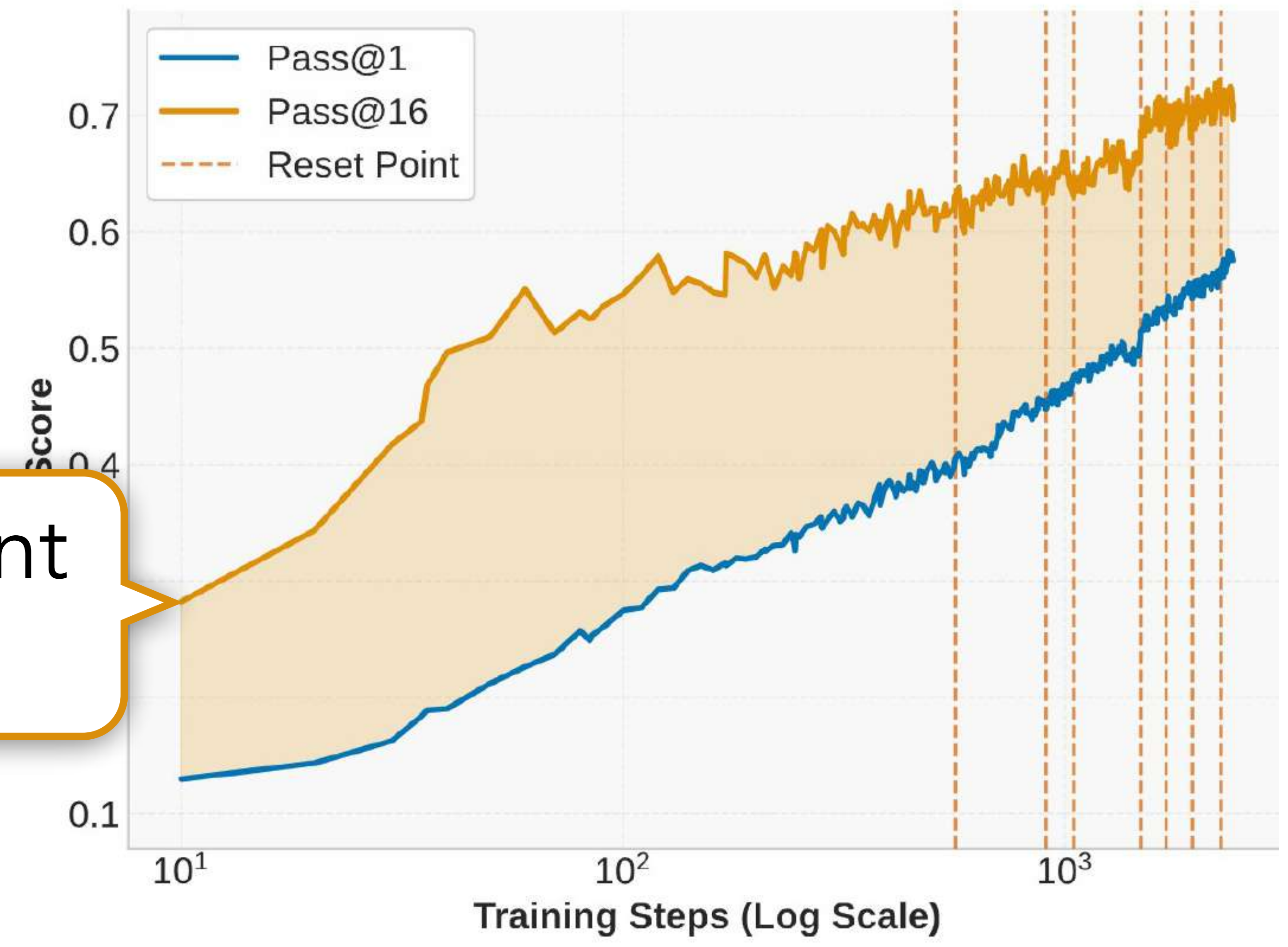






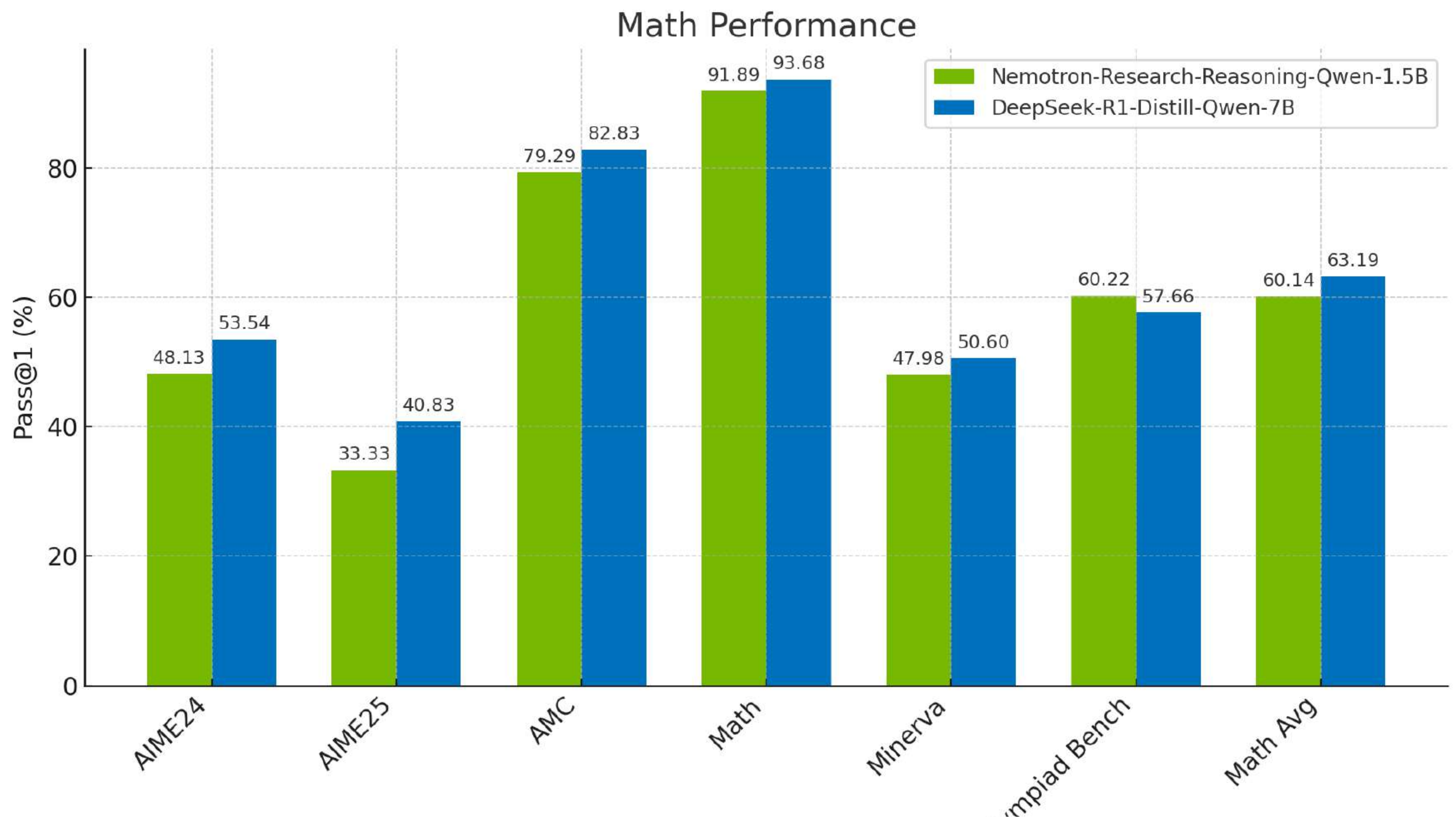




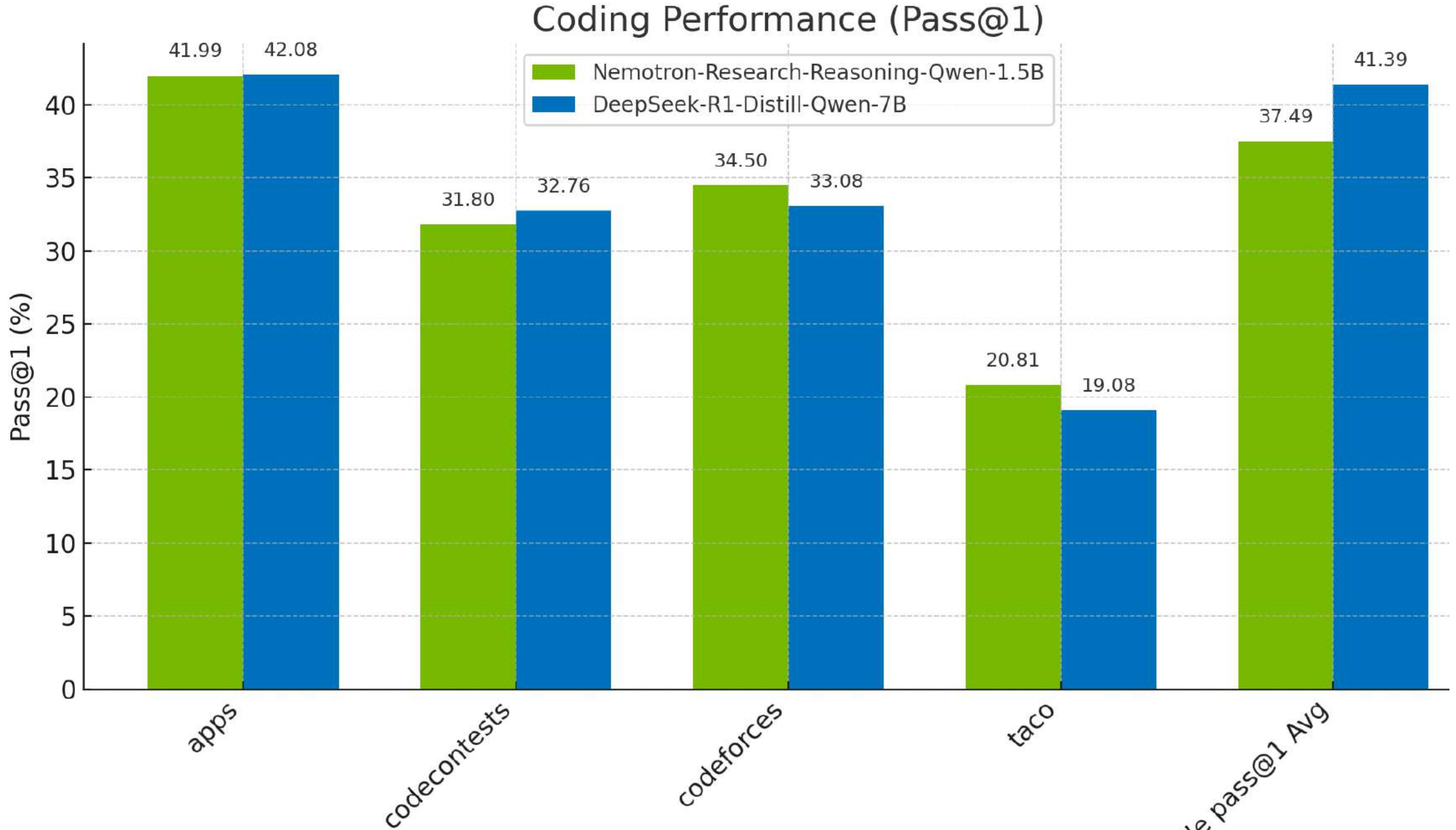


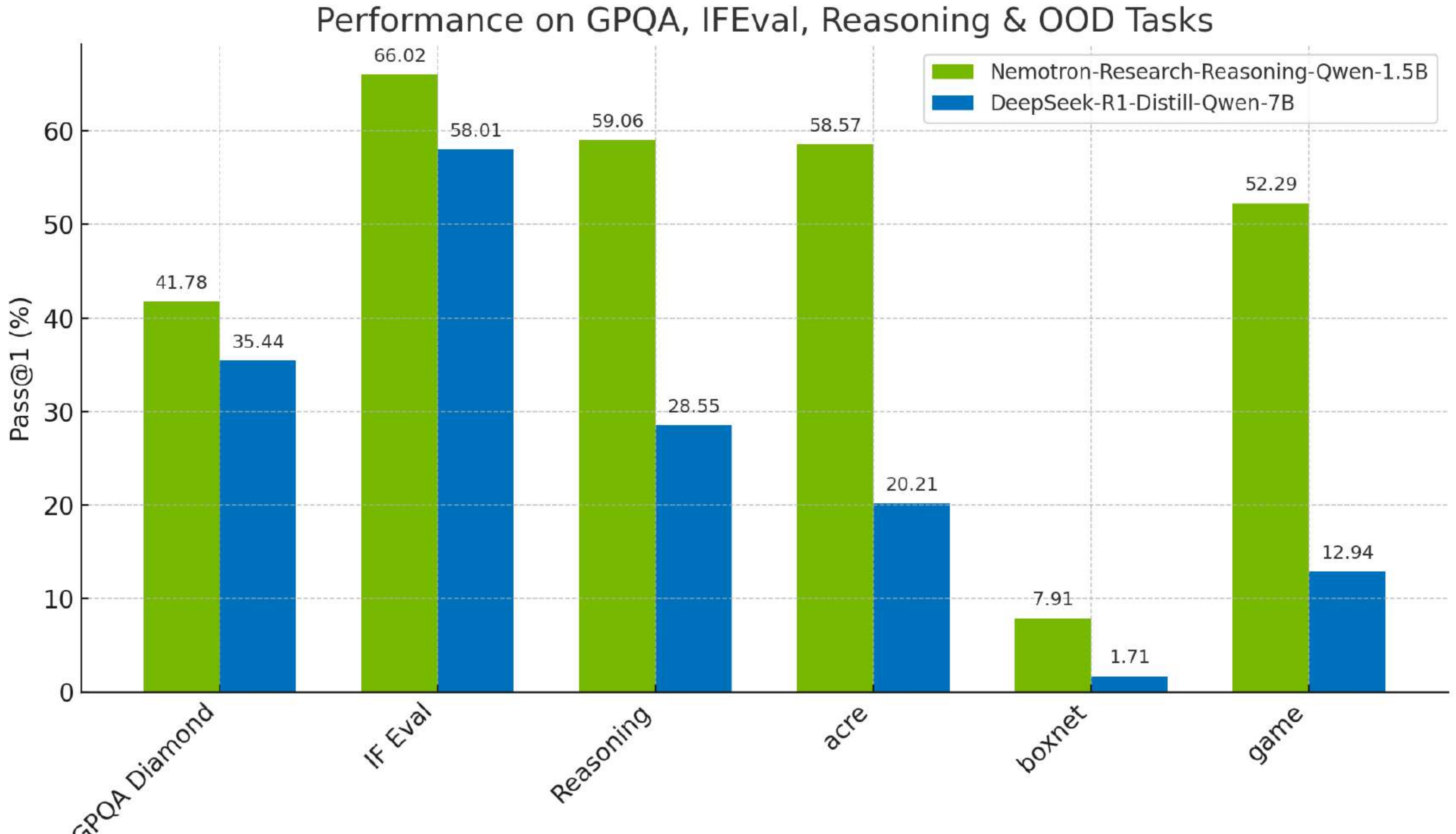
Log-linear performance improvement in both Pass@1 and Pass@16.

The SOTA 1.5B Reasoning Model

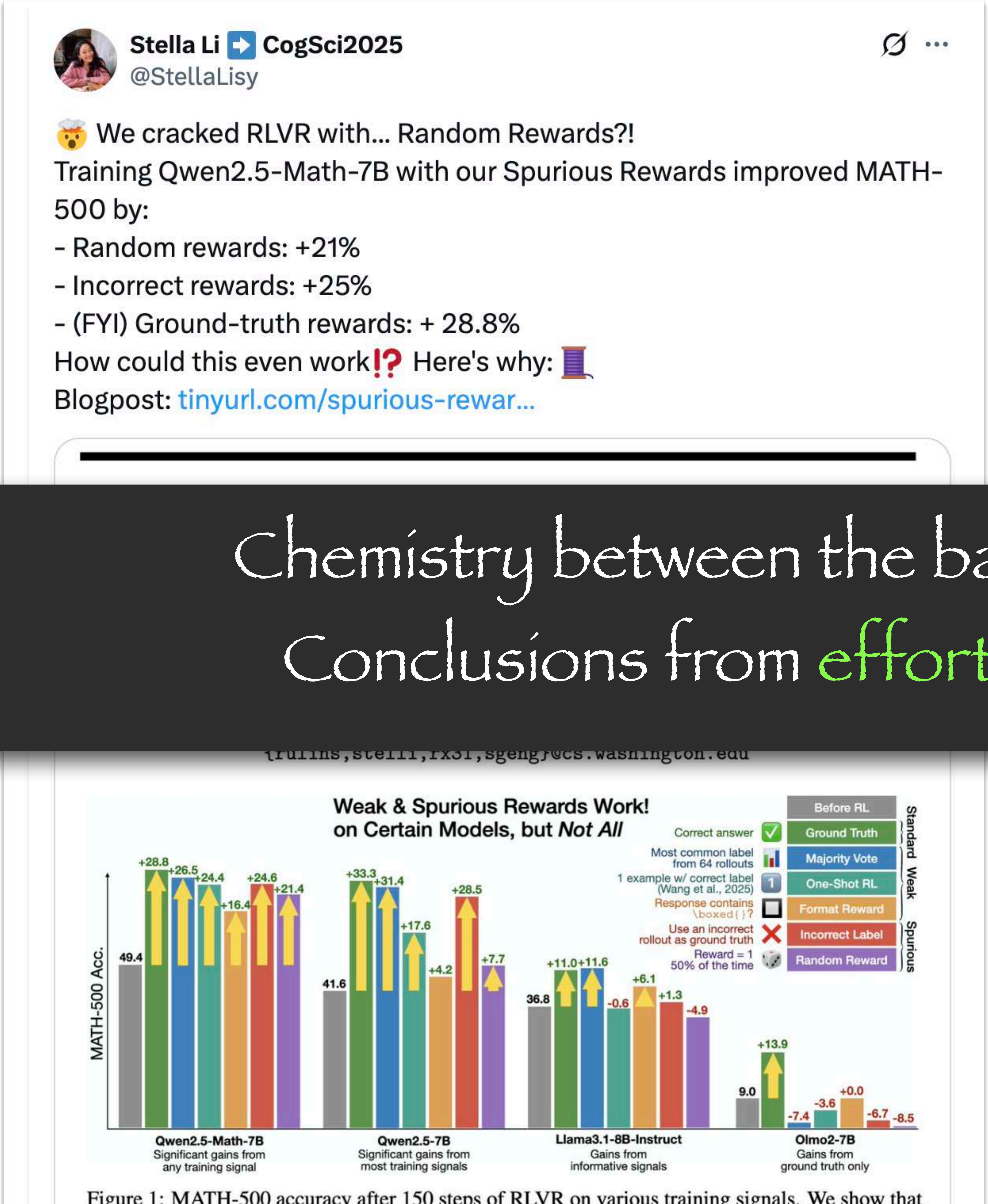


Our ProRL-ed model, **Nemotron-Reasoning -1.5B**, **achieves comparable or better performance than the 4.5× larger model, DeepSeek-R1-7B**.





Striking findings against RL



Stella Li ➡ CogSci2025
@StellaLisy

🤯 We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved MATH-500 by:
- Random rewards: +21%
- Incorrect rewards: +25%
- (FYI) Ground-truth rewards: + 28.8%

How could this even work⁉️ Here's why: 🧵
Blogpost: tinyurl.com/spurious-rewar...

{rulins,stelli,rx31,sgeng}@cs.washington.edu

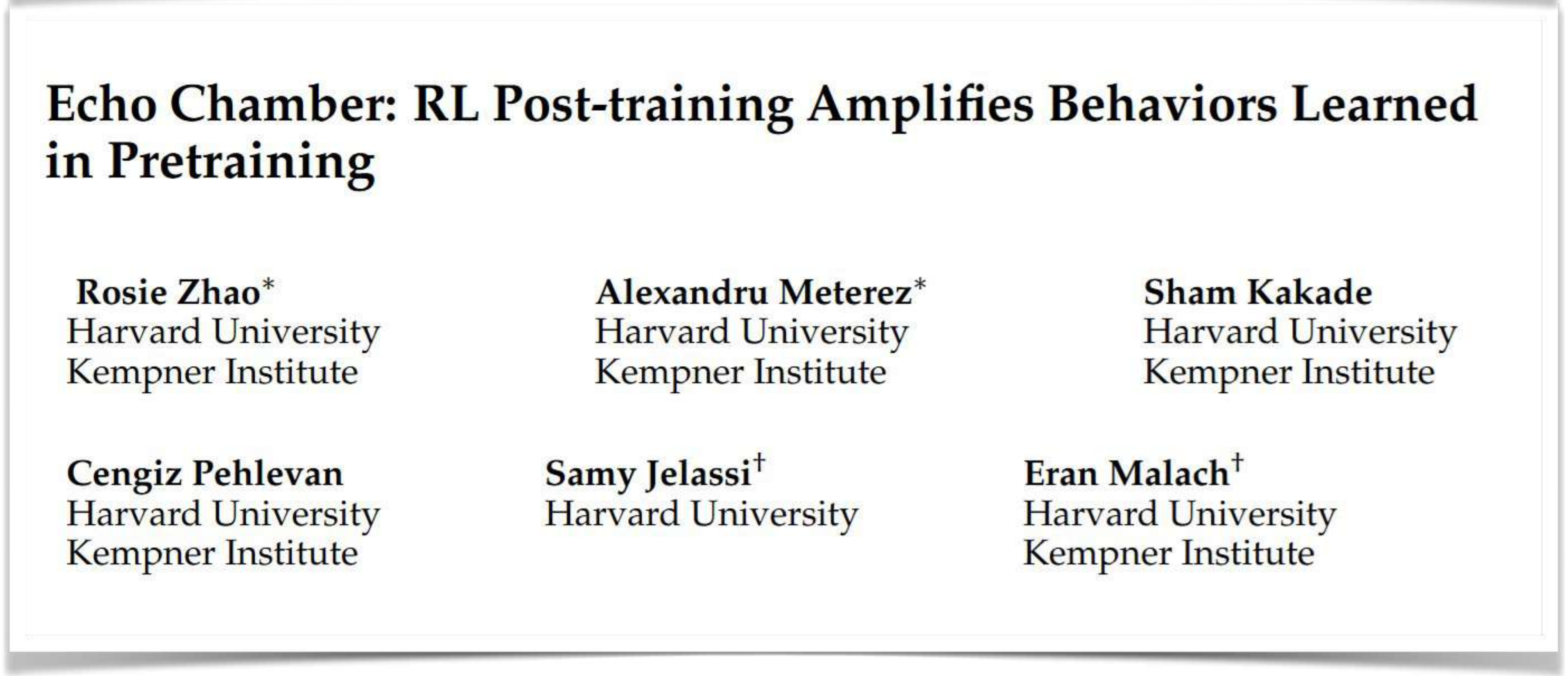
Weak & Spurious Rewards Work!
on Certain Models, but *Not All*
Before RL
Correct answer ✓ Ground Truth
Most common label from 64 rollouts Majority Vote
1 example w/ correct label (Wang et al., 2025) One-Shot RL
Response contains \boxed{}? Format Reward
Use an incorrect rollout as ground truth ✗ Incorrect Label
Reward = 1 50% of the time Random Reward
Standard
Weak
Spurious
MATH-500 Acc.
49.4
+28.8
+26.5
+24.4
+16.4
+24.6
+21.4
41.6
+33.3
+31.4
+17.6
+4.2
+28.5
+7.7
36.8
+11.0
+11.6
-0.6
+6.1
+1.3
-4.9
9.0
+13.9
-7.4
-3.6
+0.0
-6.7
-8.5
Qwen2.5-Math-7B
Significant gains from any training signal
Qwen2.5-7B
Significant gains from most training signals
Llama3.1-8B-Instruct
Gains from informative signals
Olmo2-7B
Gains from ground truth only

Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that

Chemistry between the base LLM and RL matters

Conclusions from effortless RL != effortful RL

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



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

Rosie Zhao*
Harvard University
Kempner Institute

Alexandru Meterez*
Harvard University
Kempner Institute

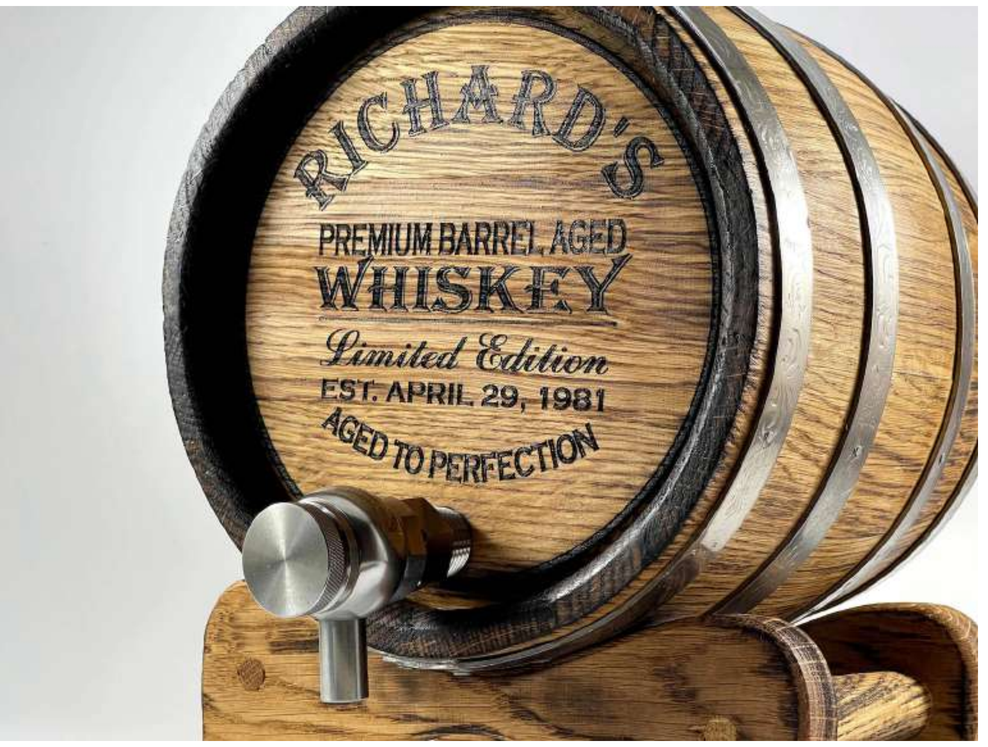
Sham Kakade
Harvard University
Kempner Institute

Cengiz Pehlevan
Harvard University
Kempner Institute

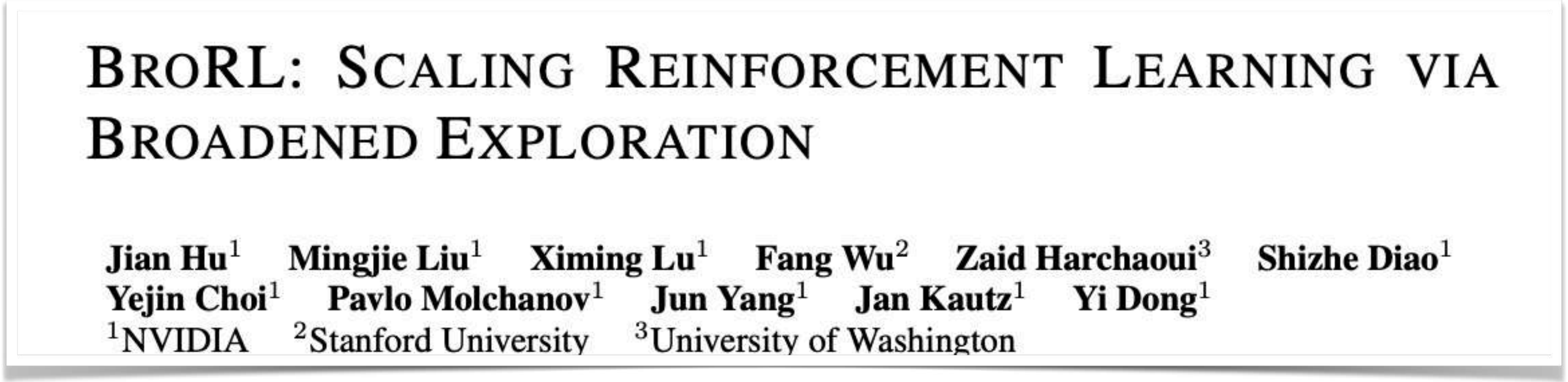
Samy Jelassi†
Harvard University

Eran Malach†
Harvard University
Kempner Institute

Conclusion (🌹 vs ☠️)



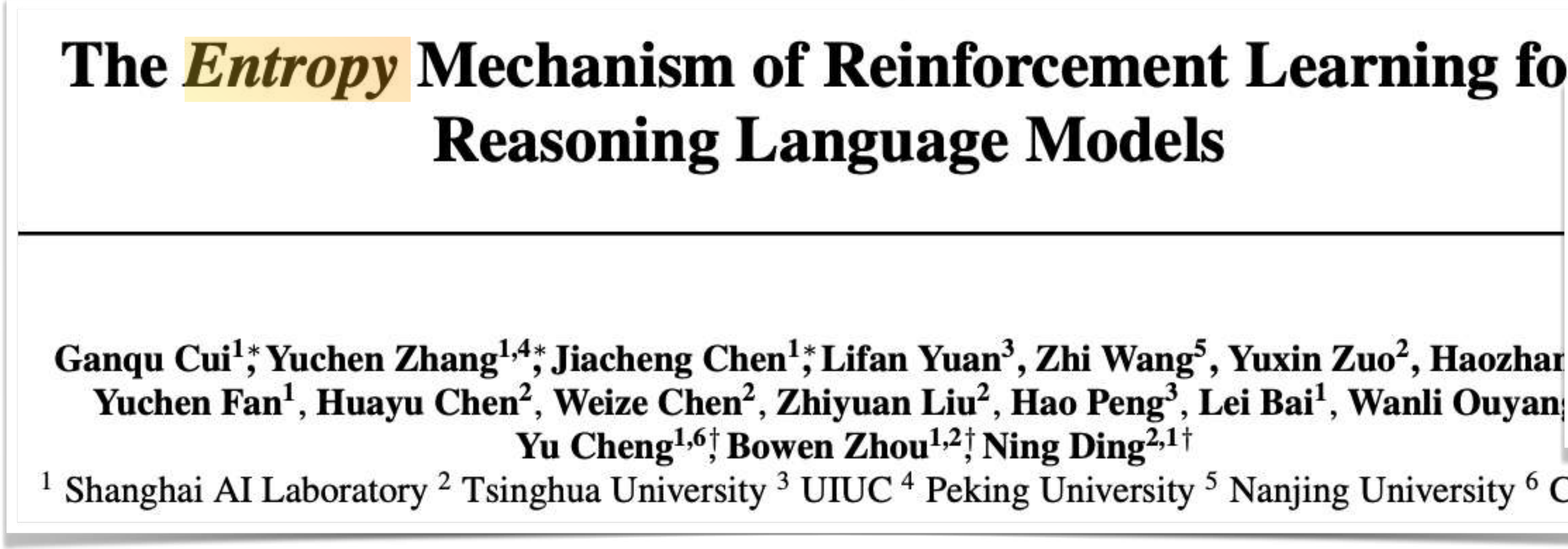
- 🌹 Conclusions from effortless RL != effortful RL
- 🌹 An effortful, prolooooooonged RL on 1.5B model can go faaaaaaaaaarrrr
- 🌹 ProRL shows that a dynamic control of entropy is critical



BRORL: SCALING REINFORCEMENT LEARNING VIA BROADENED EXPLORATION

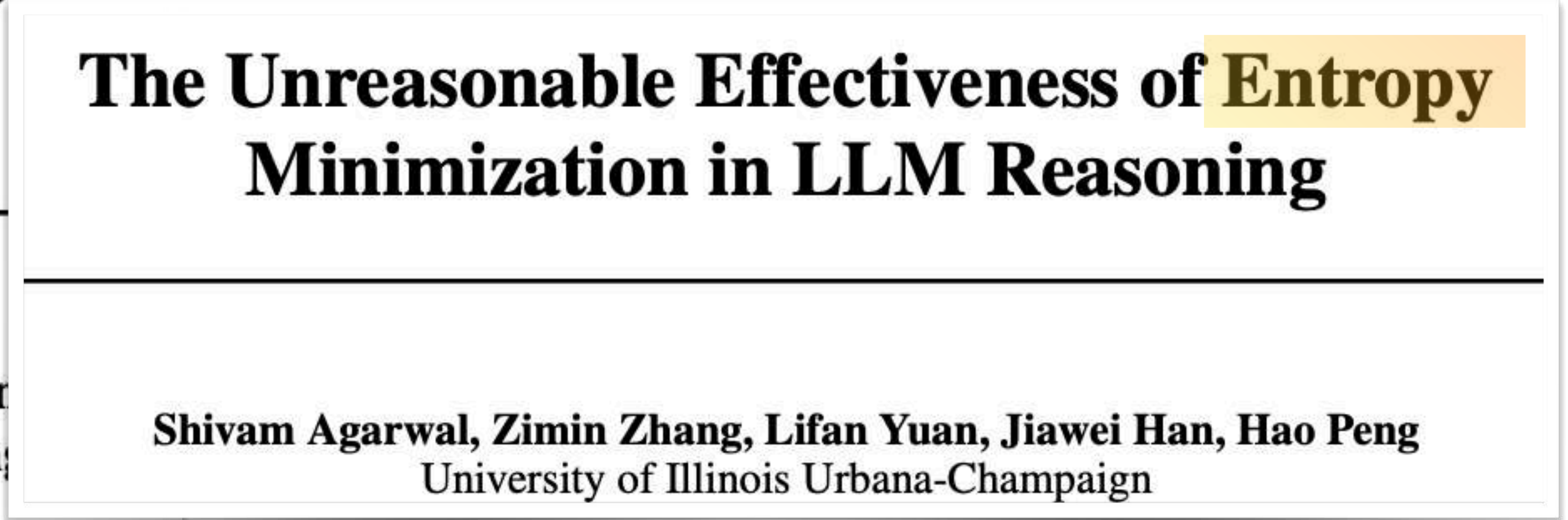
Jian Hu[1] Mingjie Liu[1] Ximing Lu[1] Fang Wu[2] Zaid Harchaoui[3] Shizhe Diao[1]
Yejin Choi[1] Pavlo Molchanov[1] Jun Yang[1] Jan Kautz[1] Yi Dong[1]
[1]NVIDIA [2]Stanford University [3]University of Washington

☠️ Can effortful RL succeed on say, GPT2?



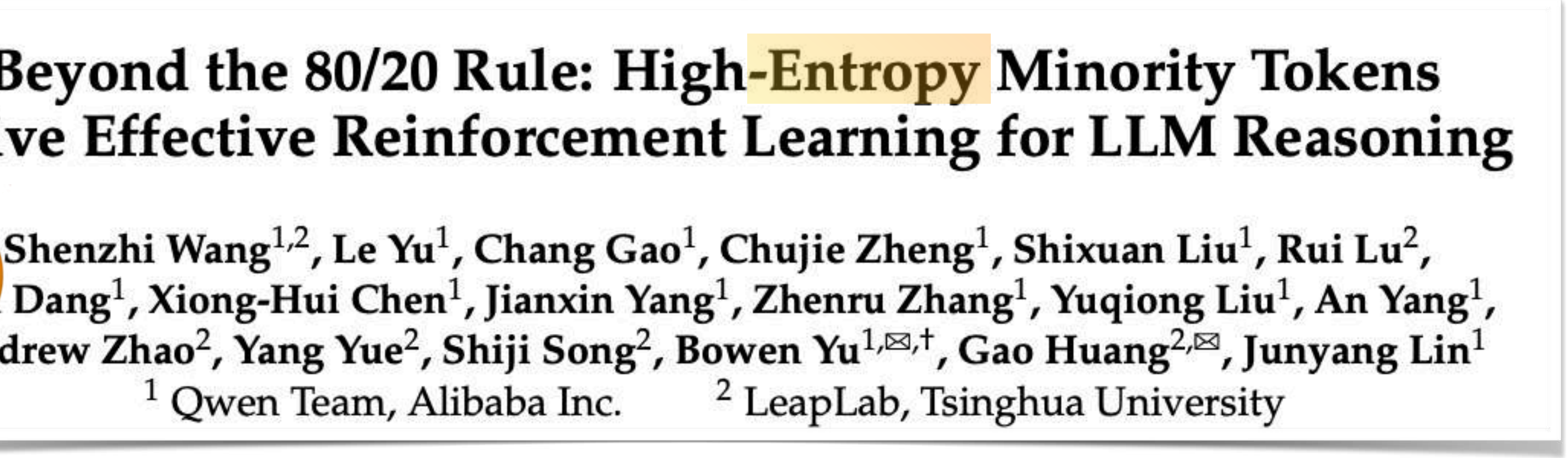
The *Entropy* Mechanism of Reinforcement Learning for Reasoning Language Models

Ganqu Cui[1]*, Yuchen Zhang[1,4]*, Jiacheng Chen[1]*, Lifan Yuan[3], Zhi Wang[5], Yuxin Zuo[2], Haozhan
Yuchen Fan[1], Huayu Chen[2], Weize Chen[2], Zhiyuan Liu[2], Hao Peng[3], Lei Bai[1], Wanli Ouyan
Yu Cheng[1,6]†, Bowen Zhou[1,2]†, Ning Ding[2,1]†
[1] Shanghai AI Laboratory [2] Tsinghua University [3] UIUC [4] Peking University [5] Nanjing University [6] CUHK



The Unreasonable Effectiveness of Entropy Minimization in LLM Reasoning

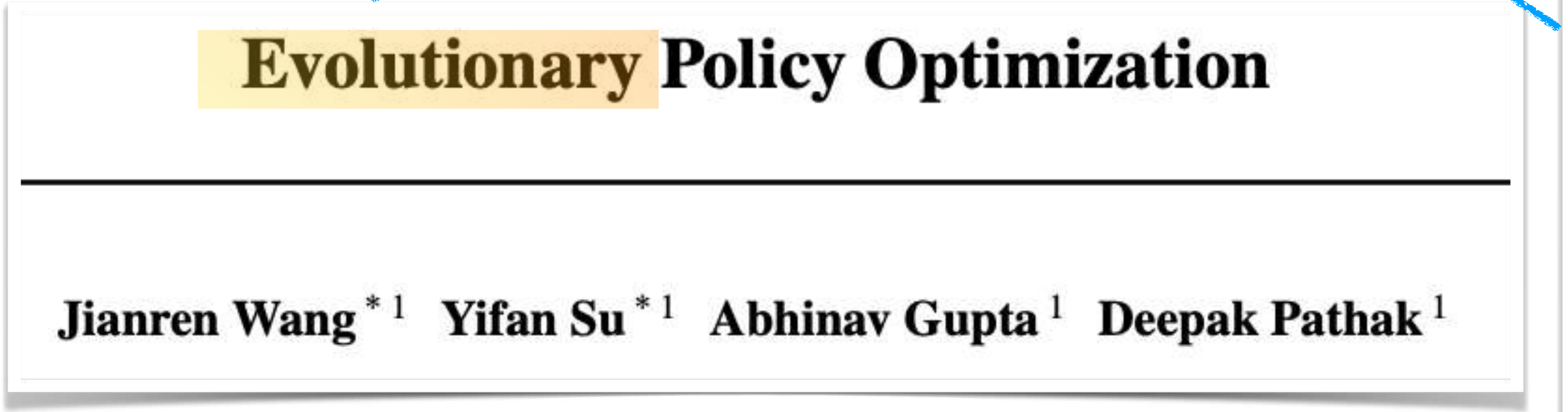
Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, Hao Peng
University of Illinois Urbana-Champaign



Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning

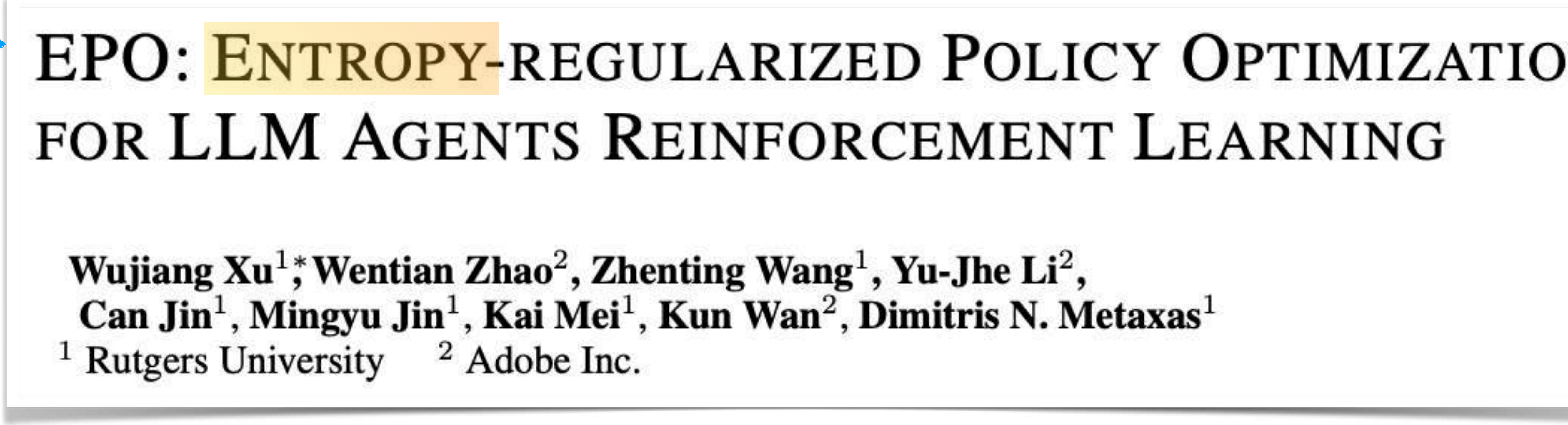
Shenzhi Wang[1,2], Le Yu[1], Chang Gao[1], Chujie Zheng[1], Shixuan Liu[1], Rui Lu[2],
Dang[1], Xiong-Hui Chen[1], Jianxin Yang[1], Zhenru Zhang[1], Yuqiong Liu[1], An Yang[1],
Andrew Zhao[2], Yang Yue[2], Shiji Song[2], Bowen Yu[1,✉,†], Gao Huang[2,✉], Junyang Lin[1]
[1] Qwen Team, Alibaba Inc. [2] LeapLab, Tsinghua University

EPO😅 EPO😅



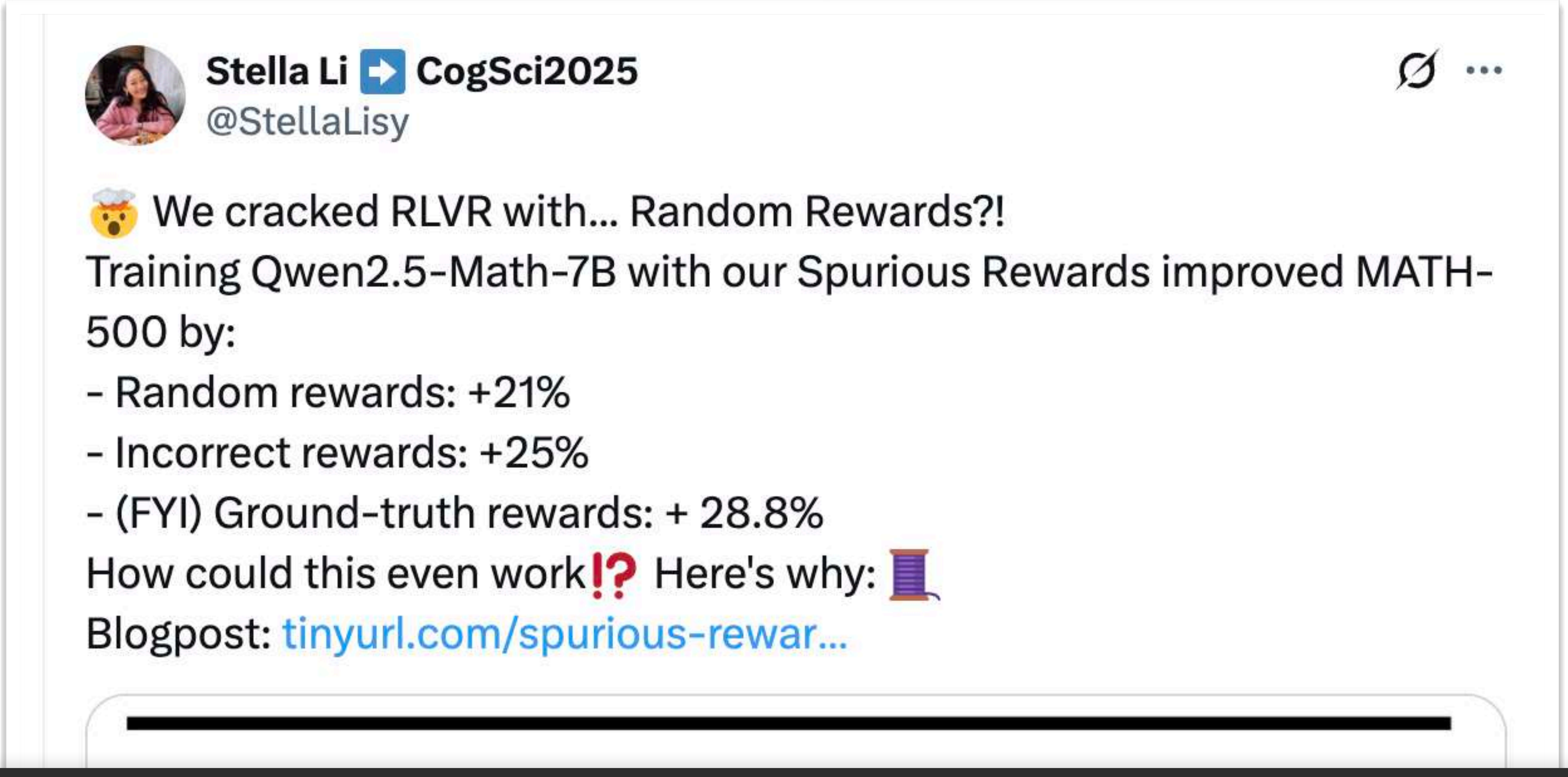
Evolutionary Policy Optimization

Jianren Wang*[1] Yifan Su*[1] Abhinav Gupta[1] Deepak Pathak[1]



EPO: ENTROPY-REGULARIZED POLICY OPTIMIZATIO FOR LLM AGENTS REINFORCEMENT LEARNING

Wujiang Xu[1]*, Wentian Zhao[2], Zhenting Wang[1], Yu-Jhe Li[2],
Can Jin[1], Mingyu Jin[1], Kai Mei[1], Kun Wan[2], Dimitris N. Metaxas[1]
[1] Rutgers University [2] Adobe Inc.



Stella Li ➡️ CogSci2025
@StellaLisy

🤯 We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved MATH-500 by:
- Random rewards: +21%
- Incorrect rewards: +25%
- (FYI) Ground-truth rewards: + 28.8%
How could this even work⁉️ Here's why: 🧵
Blogpost: tinyurl.com/spurious-rewar...

Striking findings against RL

Chemistry between the base LLM and RL matters

Conclusions from effortless RL != effortful RL

{rulins,stelli,rx31,sgeng}@cs.washington.edu

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

Conclusions from effortless SFT != effortful SFT

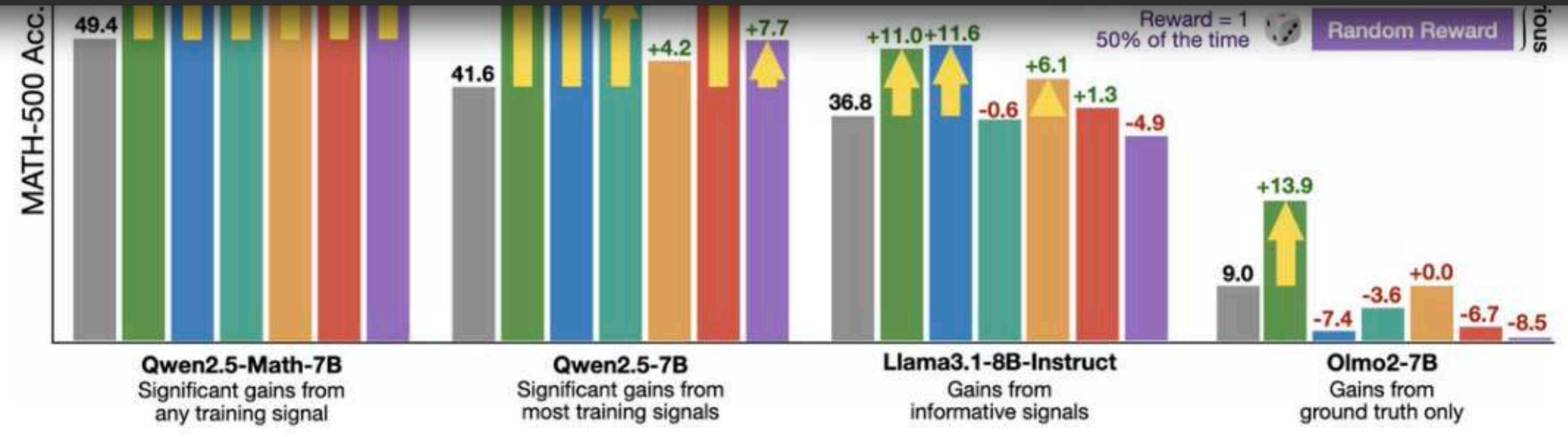
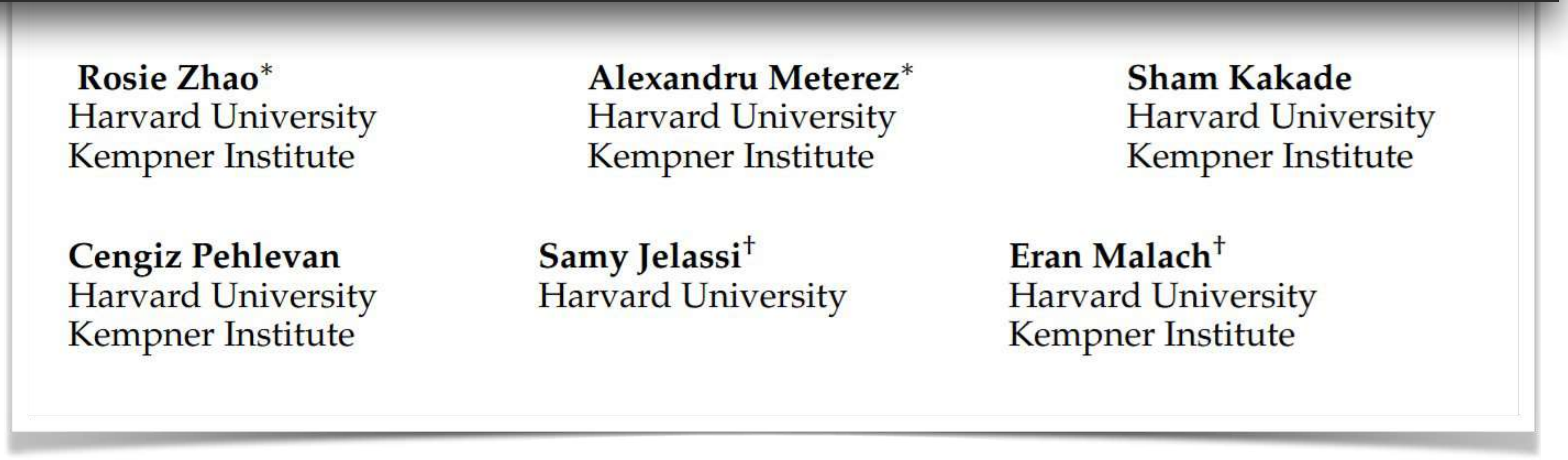


Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that



Rosie Zhao*
Harvard University
Kempner Institute

Alexandru Meterez*
Harvard University
Kempner Institute

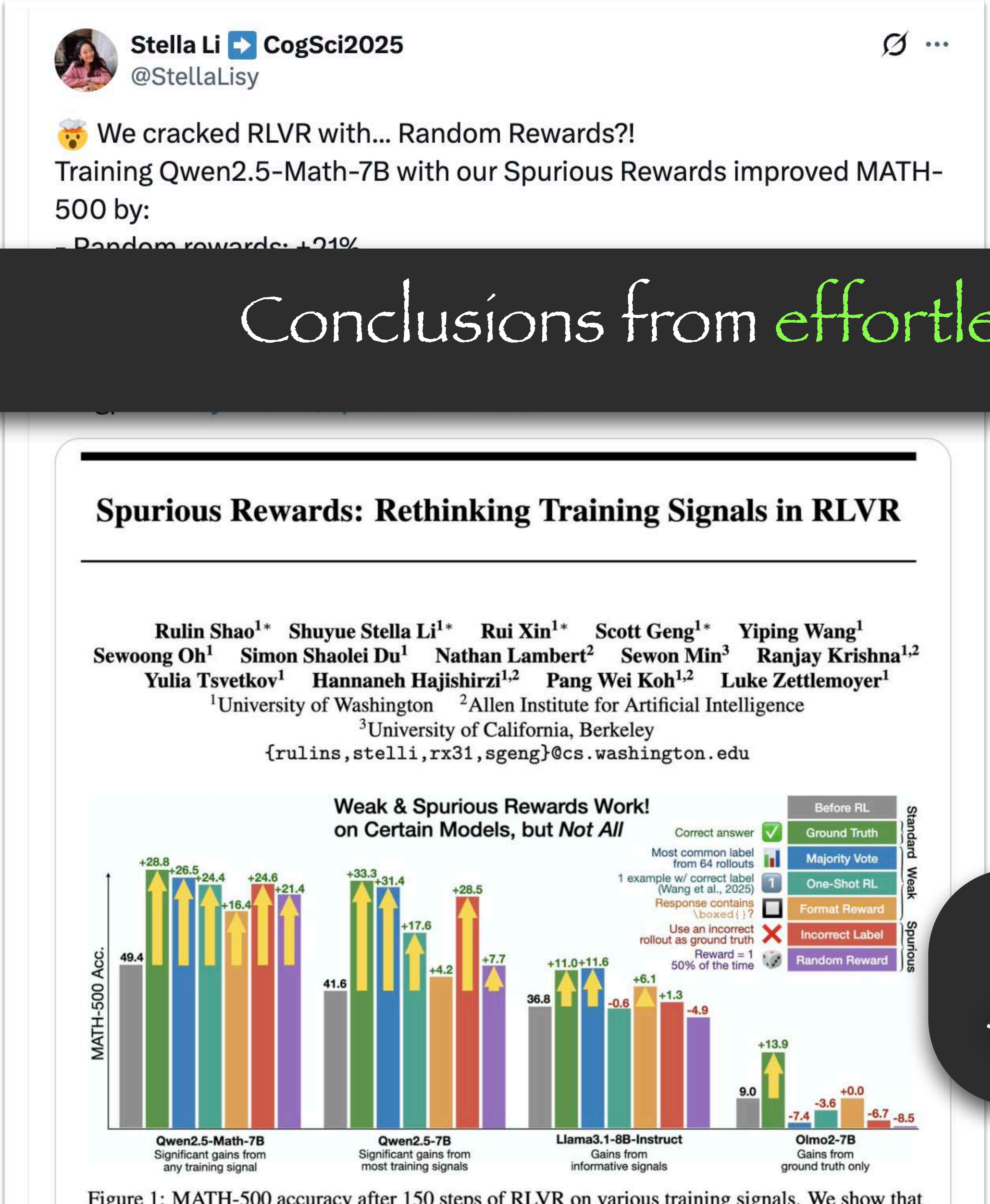
Sham Kakade
Harvard University
Kempner Institute

Cengiz Pehlevan
Harvard University
Kempner Institute

Samy Jelassi†
Harvard University

Eran Malach†
Harvard University
Kempner Institute

Striking findings against RL



Stella Li ➡ CogSci2025
@StellaLisy

🤯 We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved MATH-500 by:
- Random rewards: +21%

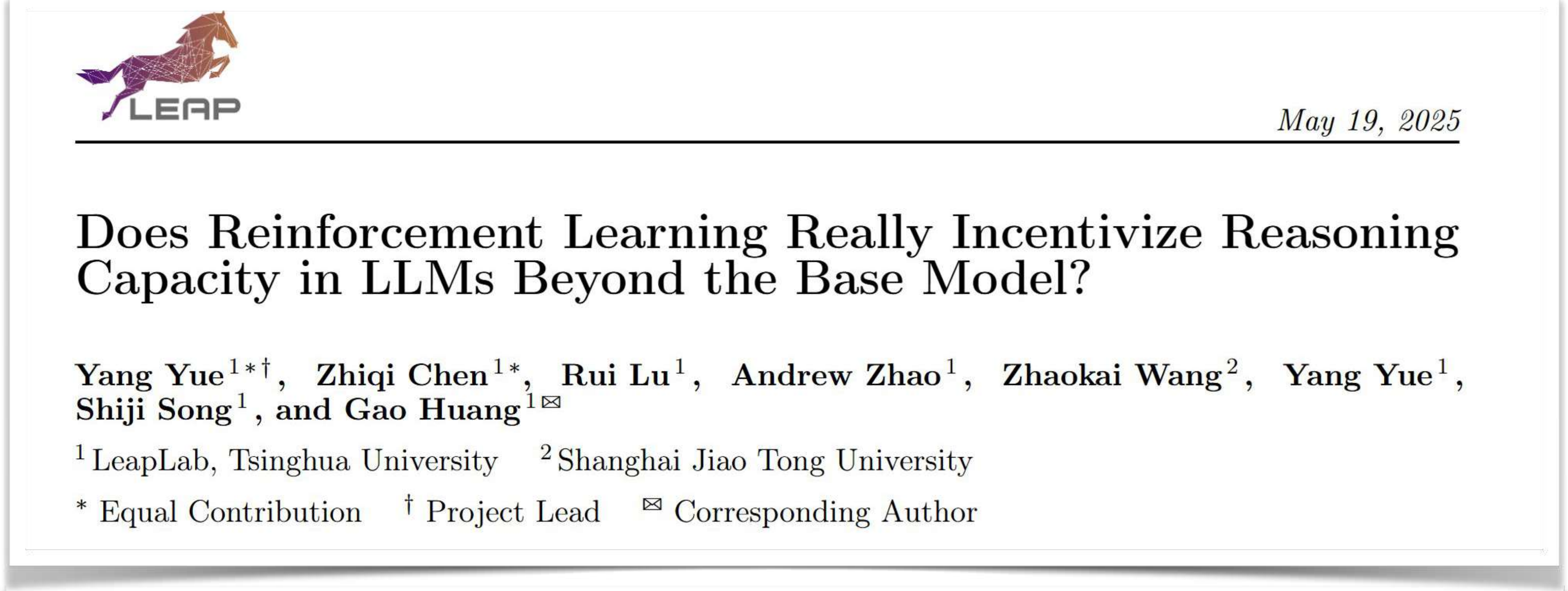
Spurious Rewards: Rethinking Training Signals in RLVR

Rulin Shao[1]* Shuyue Stella Li[1]* Rui Xin[1]* Scott Geng[1]* Yiping Wang[1]
Sewoong Oh[1] Simon Shaolei Du[1] Nathan Lambert[2] Sewon Min[3] Ranjay Krishna[1,2]
Yulia Tsvetkov[1] Hannaneh Hajishirzi[1,2] Pang Wei Koh[1,2] Luke Zettlemoyer[1]
[1]University of Washington [2]Allen Institute for Artificial Intelligence
[3]University of California, Berkeley
{rulins,stelli,rx31,sgeng}@cs.washington.edu

Weak & Spurious Rewards Work!
on Certain Models, but Not All
Correct answer ✓ Ground Truth
Most common label from 64 rollouts Majority Vote
1 example w/ correct label (Wang et al., 2025) One-Shot RL
Response contains \boxed{}? Format Reward
Use an incorrect rollout as ground truth ✗ Incorrect Label
Reward = 1 50% of the time Random Reward
Before RL
Standard
Weak
Spurious
MATH-500 Acc.
49.4
+28.8
+26.5
+24.4
+16.4
+24.6
+21.4
41.6
+33.3
+31.4
+17.6
+4.2
+28.5
+7.7
36.8
+11.0
+11.6
-0.6
+6.1
+1.3
-4.9
9.0
+13.9
-7.4
-3.6
+0.0
-6.7
-8.5
Qwen2.5-Math-7B
Significant gains from any training signal
Qwen2.5-7B
Significant gains from most training signals
Llama3.1-8B-Instruct
Gains from informative signals
Olmo2-7B
Gains from ground truth only

Figure 1: MATH-500 accuracy after 150 steps of RLVR on various training signals. We show that

Conclusions from effortless SFT != effortful SFT

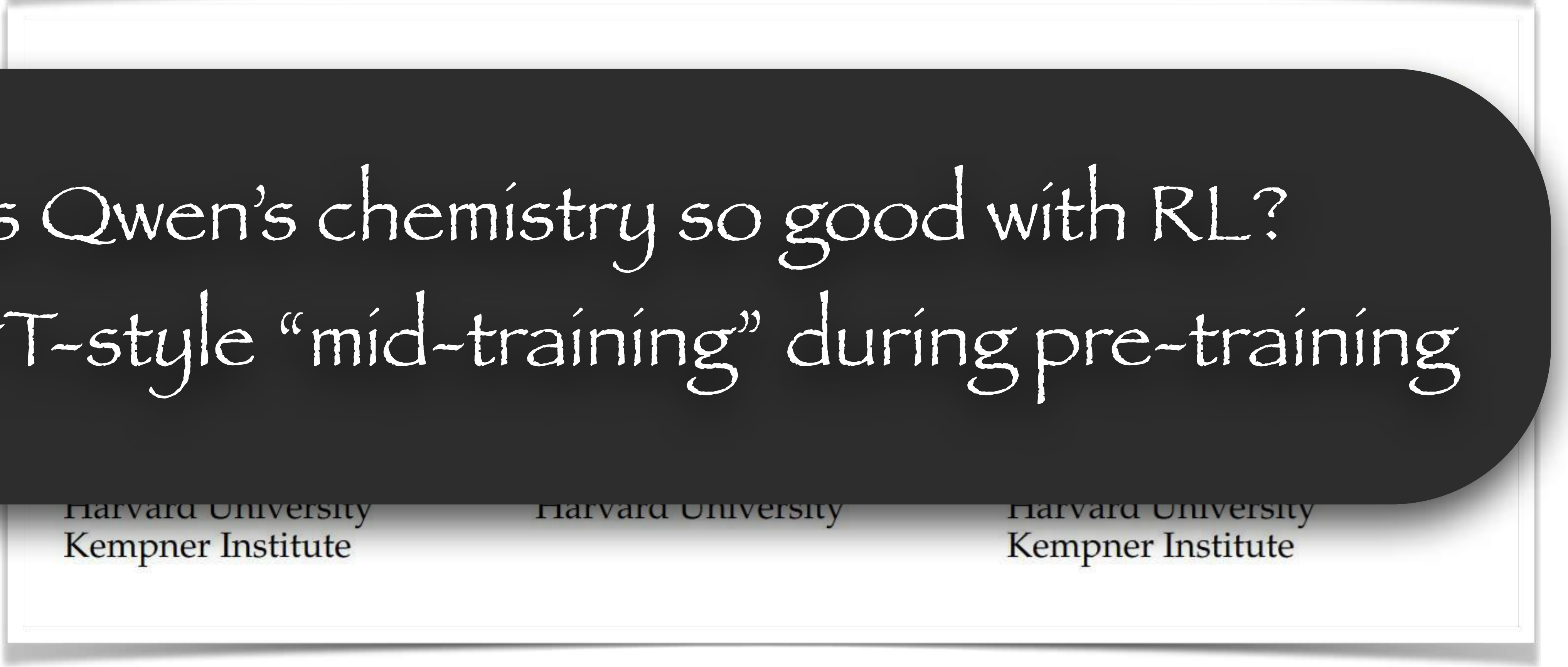


LEAP

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



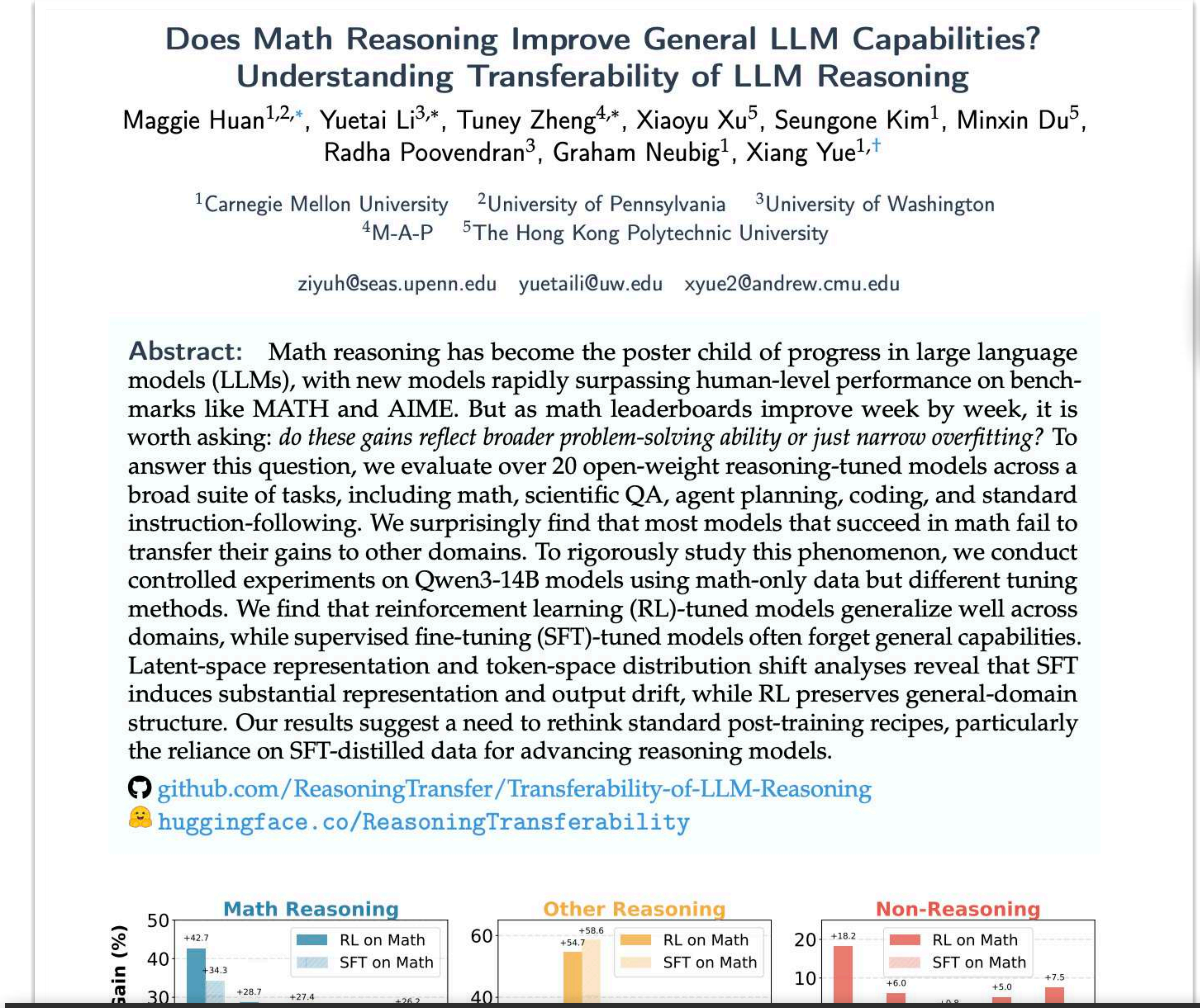
Harvard University
Kempner Institute

Harvard University

Harvard University
Kempner Institute

Why is Qwen's chemistry so good with RL?
SFT or SFT-style "mid-training" during pre-training

2025: The Rise of LRMs (as opposed to LLMs)



Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning

Maggie Huan[1,2,*], Yuetai Li[3,*], Tuney Zheng[4,*], Xiaoyu Xu[5], Seungone Kim[1], Minxin Du[5], Radha Poovendran[3], Graham Neubig[1], Xiang Yue[1,†]

[1]Carnegie Mellon University [2]University of Pennsylvania [3]University of Washington [4]M-A-P [5]The Hong Kong Polytechnic University

ziyuh@seas.upenn.edu yuetaili@uw.edu xyue2@andrew.cmu.edu

Abstract: Math reasoning has become the poster child of progress in large language models (LLMs), with new models rapidly surpassing human-level performance on benchmarks like MATH and AIME. But as math leaderboards improve week by week, it is worth asking: *do these gains reflect broader problem-solving ability or just narrow overfitting?* To answer this question, we evaluate over 20 open-weight reasoning-tuned models across a broad suite of tasks, including math, scientific QA, agent planning, coding, and standard instruction-following. We surprisingly find that most models that succeed in math fail to transfer their gains to other domains. To rigorously study this phenomenon, we conduct controlled experiments on Qwen3-14B models using math-only data but different tuning methods. We find that reinforcement learning (RL)-tuned models generalize well across domains, while supervised fine-tuning (SFT)-tuned models often forget general capabilities. Latent-space representation and token-space distribution shift analyses reveal that SFT induces substantial representation and output drift, while RL preserves general-domain structure. Our results suggest a need to rethink standard post-training recipes, particularly the reliance on SFT-distilled data for advancing reasoning models.

github.com/ReasoningTransfer/Transferability-of-LLM-Reasoning

huggingface.co/ReasoningTransferability

Then, SFT cannot generalize?

It depends…

SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training

Tianzhe Chu ♠ * Yuexiang Zhai ♥ ♣ * Jihan Yang ♦ Shengbang Tong ♦
Saining Xie ♣ ♦ Dale Schuurmans ♣ ♛ Quoc V. Le ♣ Sergey Levine ♥ Yi Ma ♠ ♥

Conclusions from effortless SFT != effortful SFT

In this talk:

===

ProRL: Prolonged RL

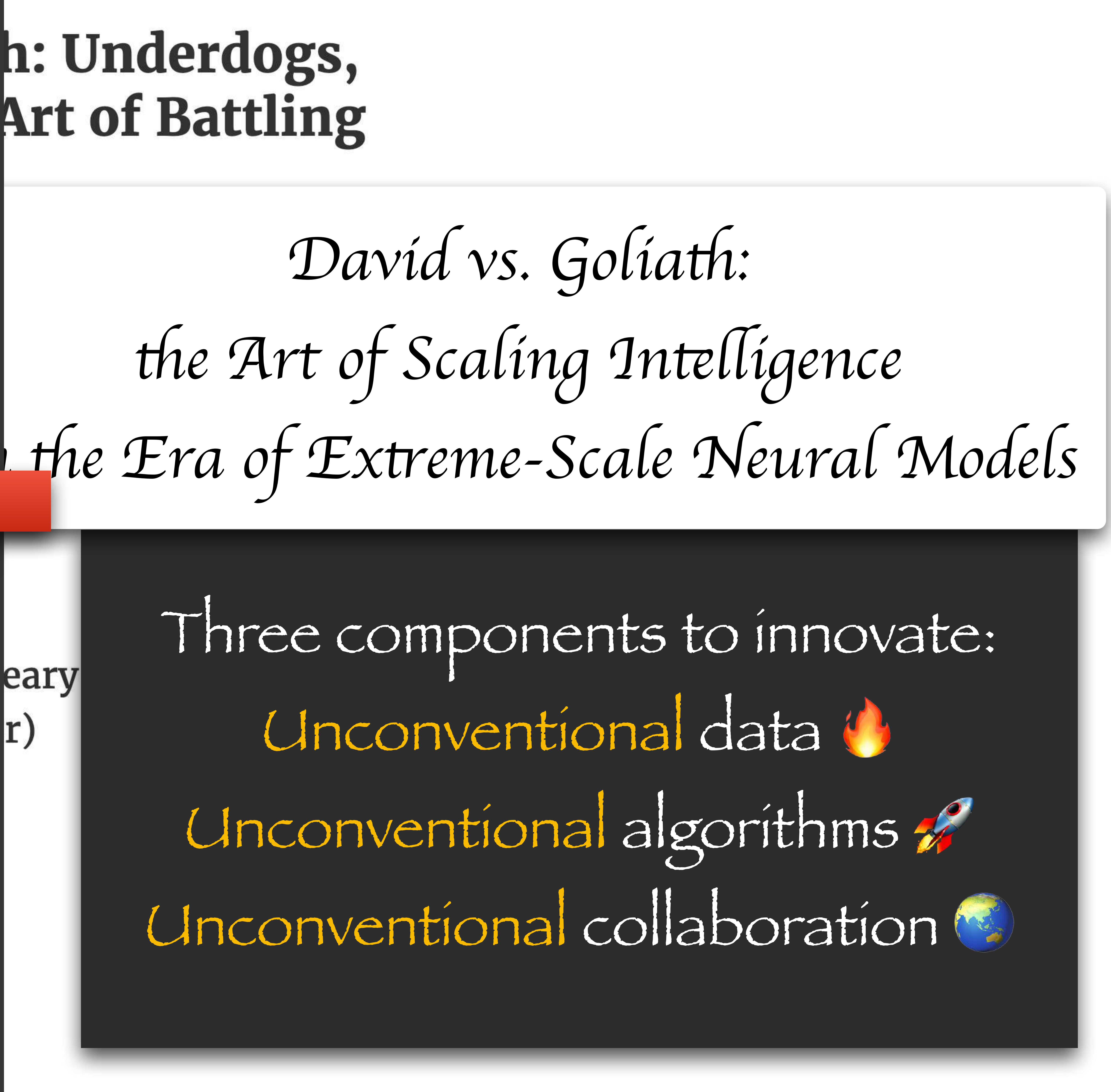
Prismatic Synthesis

RL as Pretraining

===

"Smaller but Better"

"Algorithms for the Win"



h: Underdogs,
Art of Battling

eary
r)

David vs. Goliath:
the Art of Scaling Intelligence
the Era of Extreme-Scale Neural Models

Three components to innovate:

Unconventional data 🔥

Unconventional algorithms 🚀

Unconventional collaboration 🌏

Prismatic Synthesis:

Gradient-based Data Diversification Boosts Generalization in LLM Reasoning

— NeurIPS 2025 —

Jaehun Jung

Seungju Han

Ximing Lu

Skyler Hallinan

David Acuna

Shrimai Prabhumoye

Mostafa Patwary

Mohammad Shoeybi

Bryan Catanzaro

Yejin Choi

The Era of Brute-Force Scaling is Over

The Era of Smart Scaling Begins

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

-- Ilya Sutskever from his test time award talk at NeurIPS 2024

Synthetic Data to the Rescue!

Already used in both mid / post-training:

Qwen

Microsoft

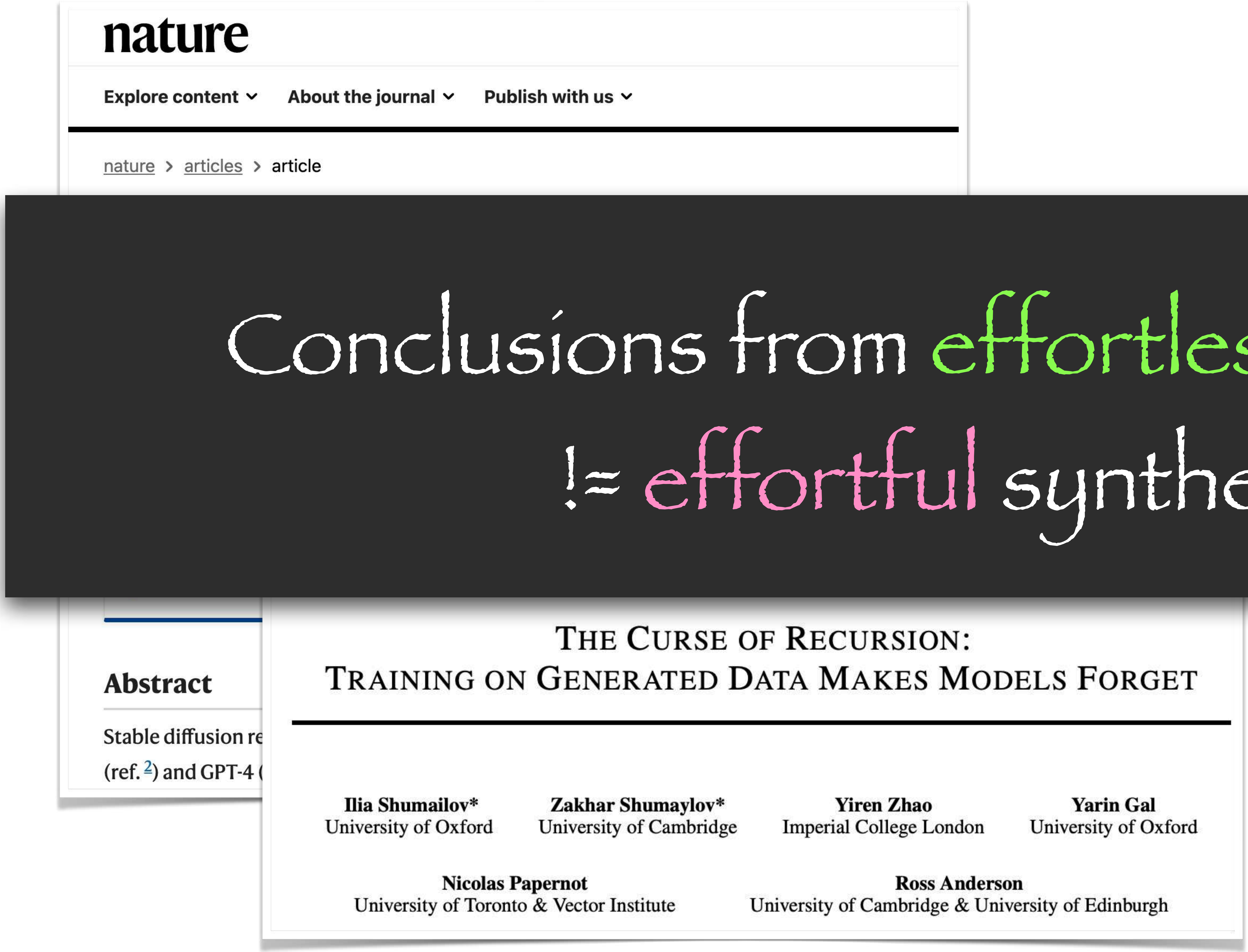
Phi-4

Sometimes, for the entirety of LLM training…

Common practice & assumption

(1) rely on the largest / strongest teacher models

Obvious Concerns: Mode Collapse



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Abstract

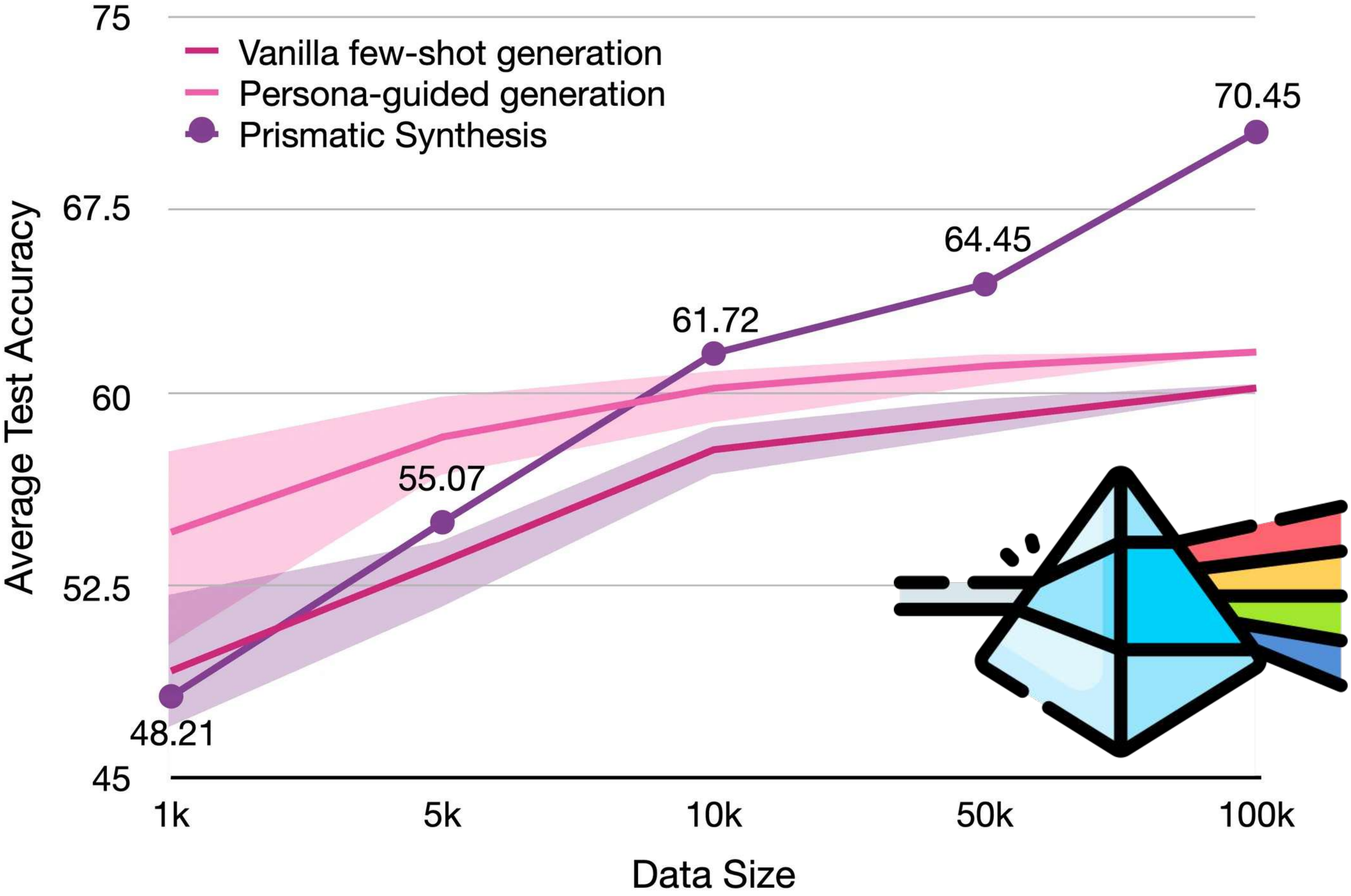
Stable diffusion re

(ref. [2]) and GPT-4 (

THE CURSE OF RECURSION:
TRAINING ON GENERATED DATA MAKES MODELS FORGET

Ilia Shumailov* University of Oxford
Zakhar Shumaylov* University of Cambridge
Yiren Zhao Imperial College London
Yarin Gal University of Oxford
Nicolas Papernot University of Toronto & Vector Institute
Ross Anderson University of Cambridge & University of Edinburgh

Conclusions from effortless synthetic data
!= effortful synthetic data

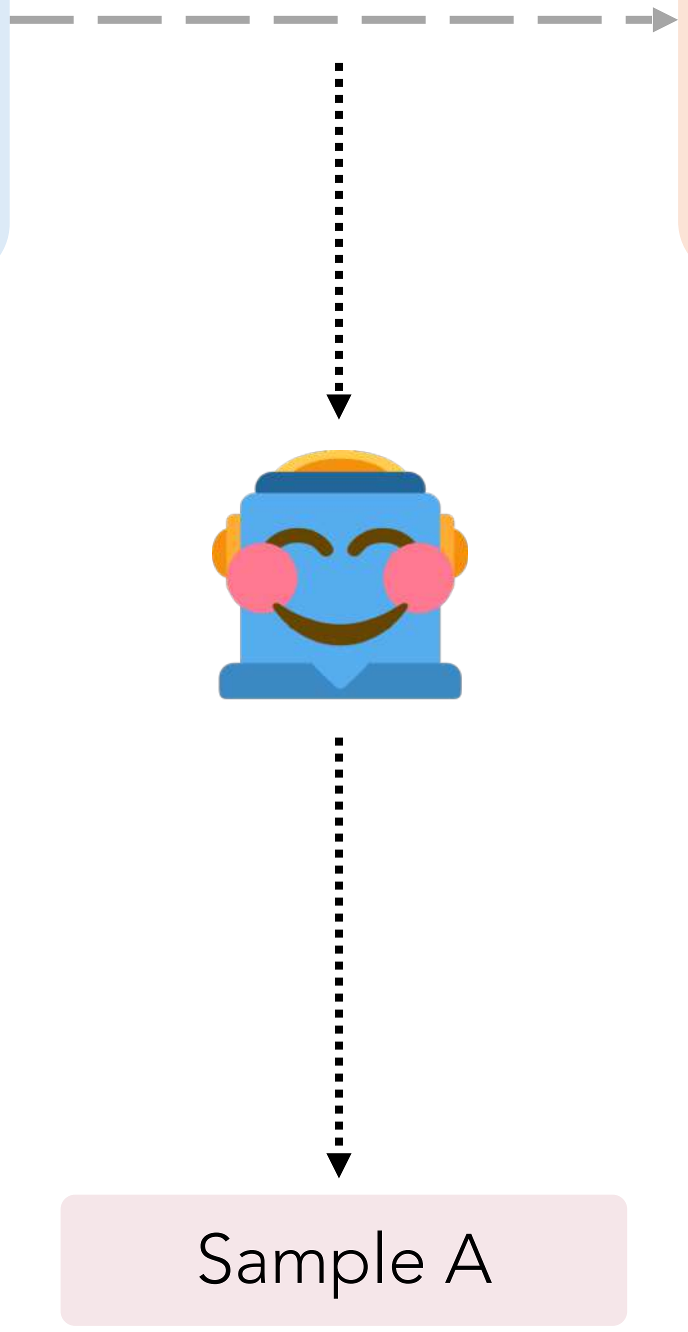


Vanilla few-shot generation
Persona-guided generation
Prismatic Synthesis
75
67.5
60
52.5
45
Average Test Accuracy
70.45
64.45
61.72
55.07
48.21
1k
5k
10k
50k
100k
Data Size

Gradient as Data Representation

Input x
Bill walks 0.5 mile south, then 0.75 mile east, and finally 0.5 mile south. How many miles is he, in a direct line, from his starting point?

Output $\mathbf{y}$
Bill walked 0.5+0.5 = 1.0 miles south total and walked 0.75 miles east in total. Therefore, using Pythagorean… the answer is 1.25 miles.



Sample A

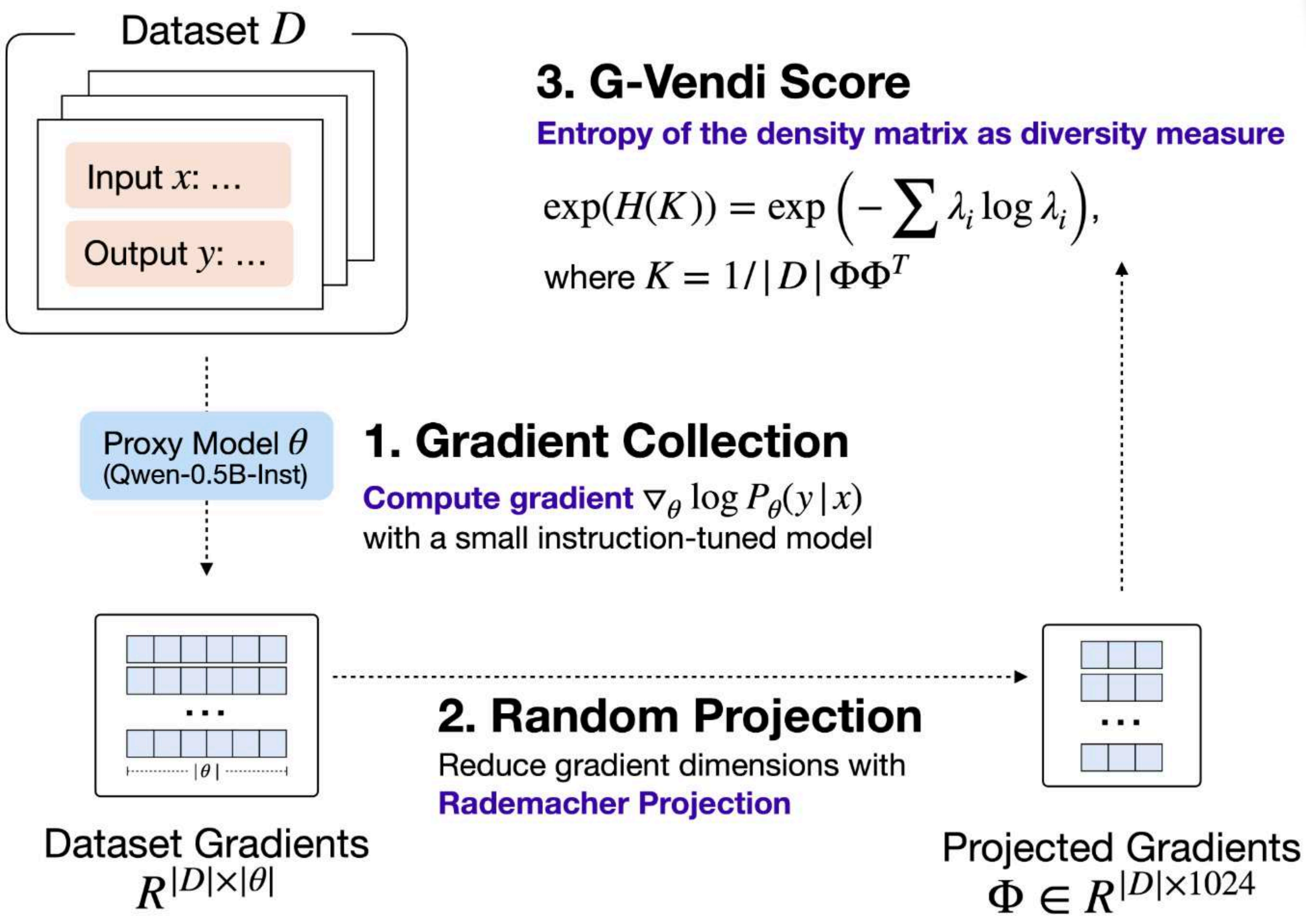
Use a small reference model to compute $\nabla_\theta P(\mathbf{y}|\mathbf{x})$

Rademacher Projection *for dimensionality reduction*

Gradient encodes the mapping from x to y;
which naturally represents
"reasoning"
between input and output!

1024-dim representation of data point

G-Vendi Score



Dataset D

Input x: …

Output y: …

Proxy Model θ (Qwen-0.5B-Inst)

1. Gradient Collection

Compute gradient $\nabla_\theta \log P_\theta(y \mid x)$ with a small instruction-tuned model

Dataset Gradients $R^{|D| \times |\theta|}$

2. Random Projection

Reduce gradient dimensions with **Rademacher Projection**

Projected Gradients $\Phi \in R^{|D| \times 1024}$

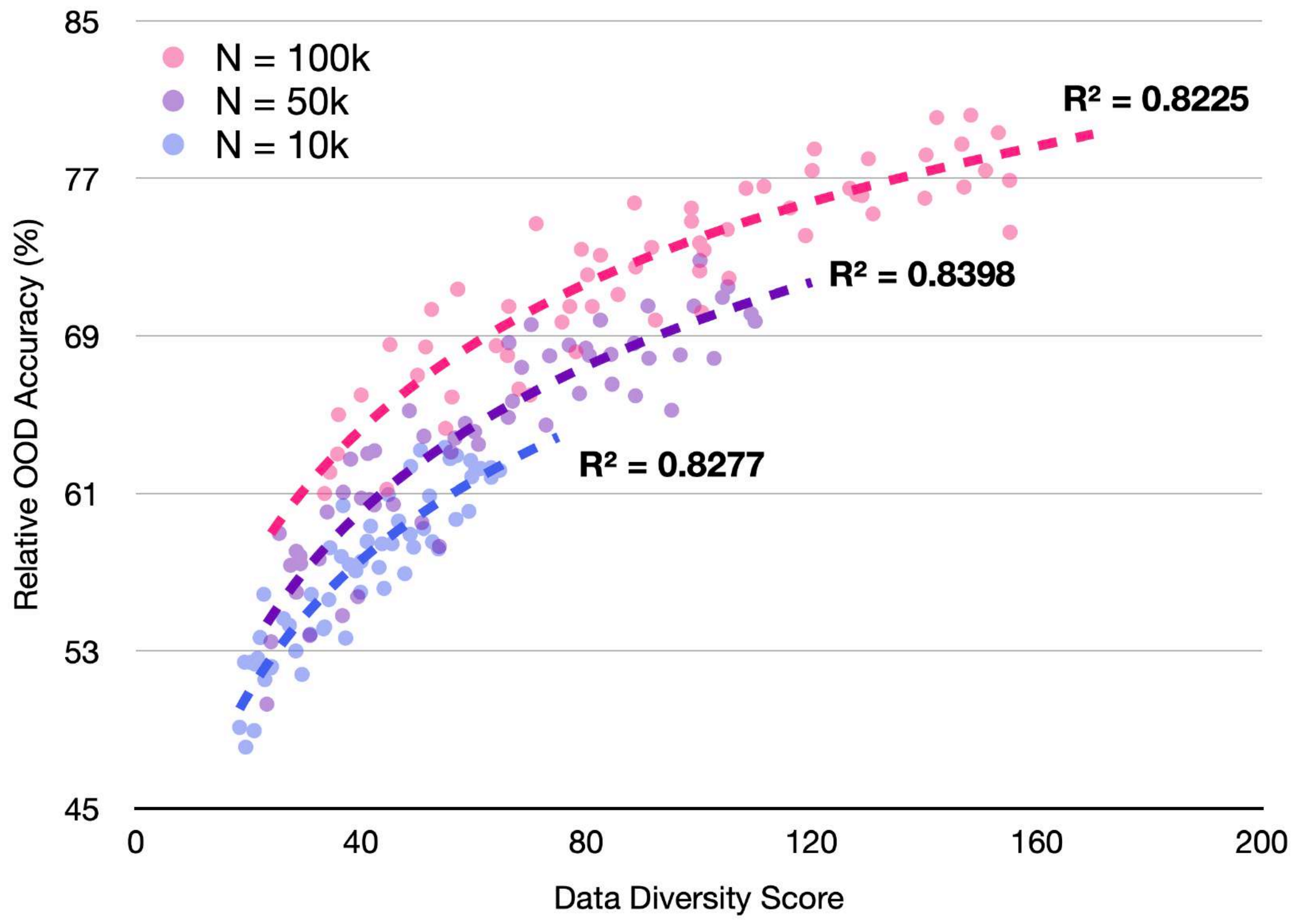
3. G-Vendi Score

Entropy of the density matrix as diversity measure

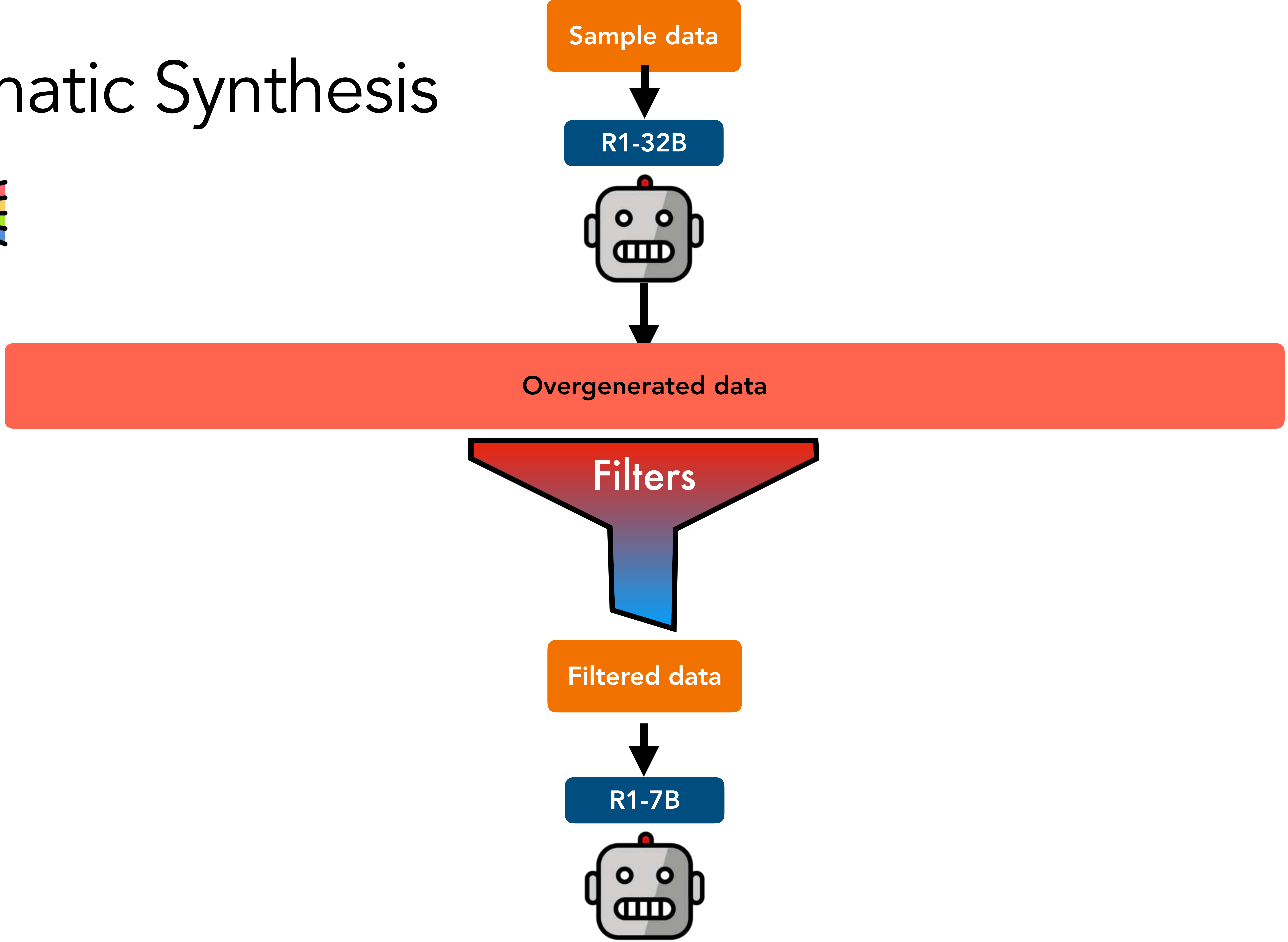
$$\exp(H(K)) = \exp\left(-\sum \lambda_i \log \lambda_i\right),$$

where $K = 1/|D|\Phi\Phi^T$

G-vendi score predicts generalization (i.e., out-of-distribution generalization)!



Prismatic Synthesis



Sample data

R1-32B

Overgenerated data

Filters

Filtered data

R1-7B

Sample data

R1-32B

Overgenerated data

Filters

Filters

Filtered data

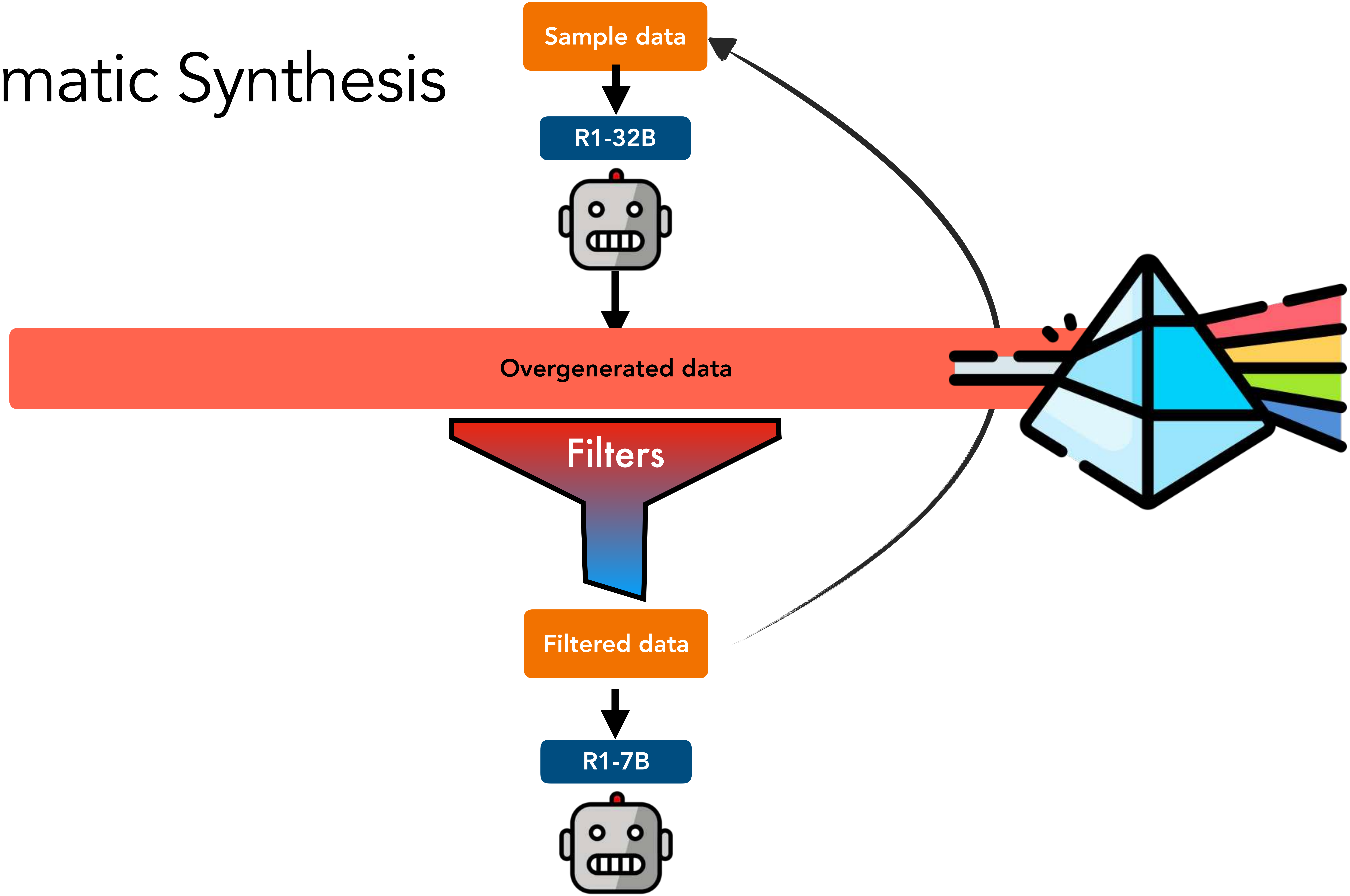
R1-7B

Gradient-based diversity filter

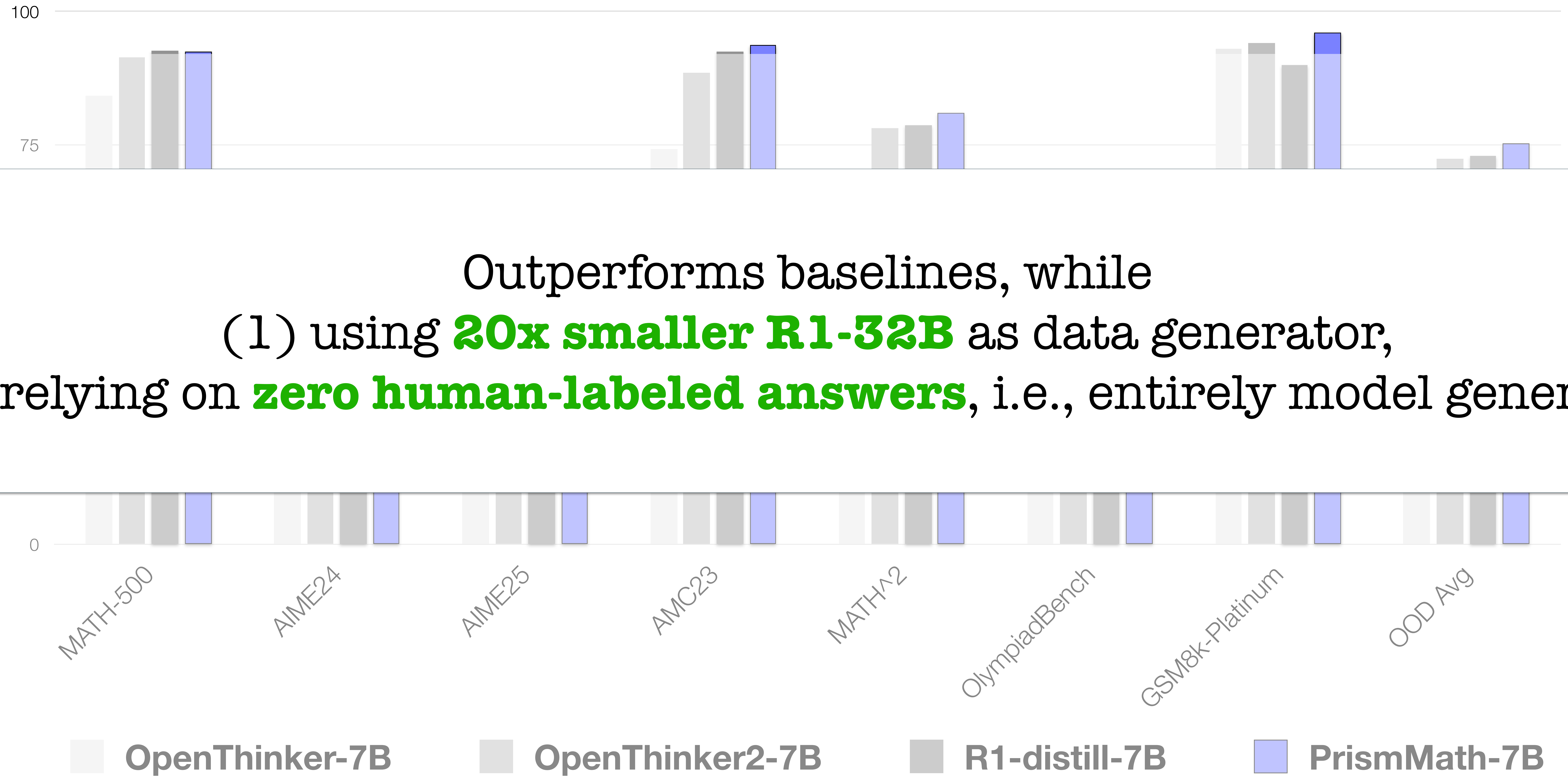
Quality filter

Discard samples based on majority votes

Prismatic Synthesis



Prismatic Synthesis improves long-CoT reasoning



Outperforms baselines, while
(1) using **20x smaller R1-32B** as data generator,
(2) relying on **zero human-labeled answers**, i.e., entirely model generated!

Concluding Thoughts

- Reasoning requires data that transcends the internet data
- Synthetic data to the rescue
- RL can be viewed as a form of (implicit) synthetic data generation via exploration
- RL or not, existing methods often lack the bird eye view on the overall diversity
- Systematic diversification (+ quality check) can make a big difference, even to overcome 20x size difference of the R1 teacher models

In this talk:

===

ProRL: Prolonged RL

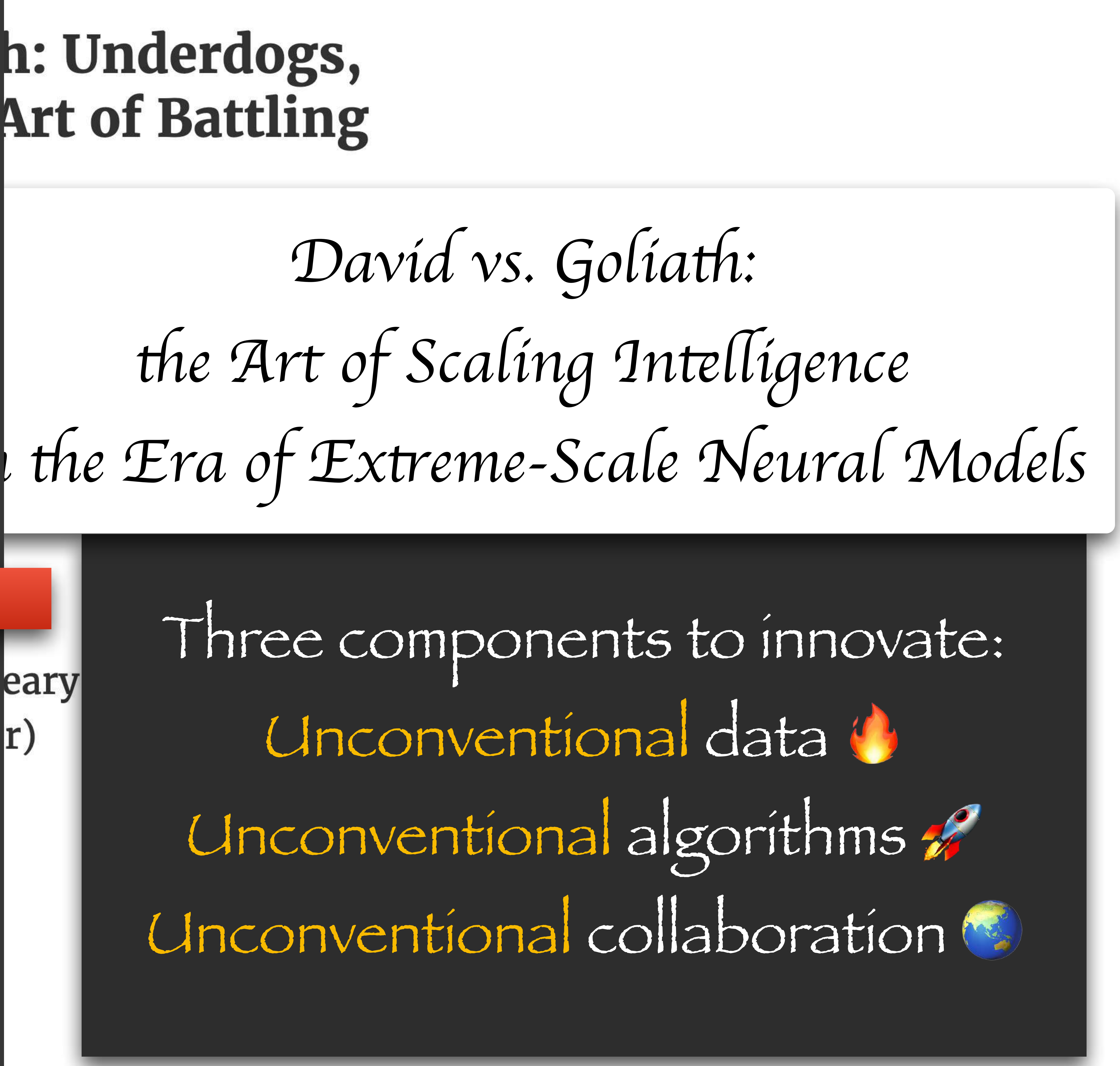
Prismatic Synthesis

RL as Pretraining

===

"Smaller but Better"

"Algorithms for the Win"



h: Underdogs,
Art of Battling

eary
r)

David vs. Goliath:
the Art of Scaling Intelligence
the Era of Extreme-Scale Neural Models

Three components to innovate:

Unconventional data 🔥

Unconventional algorithms 🚀

Unconventional collaboration 🌏

RLP

Reinforcement as a Pretraining Objective

— ICLR 2026—

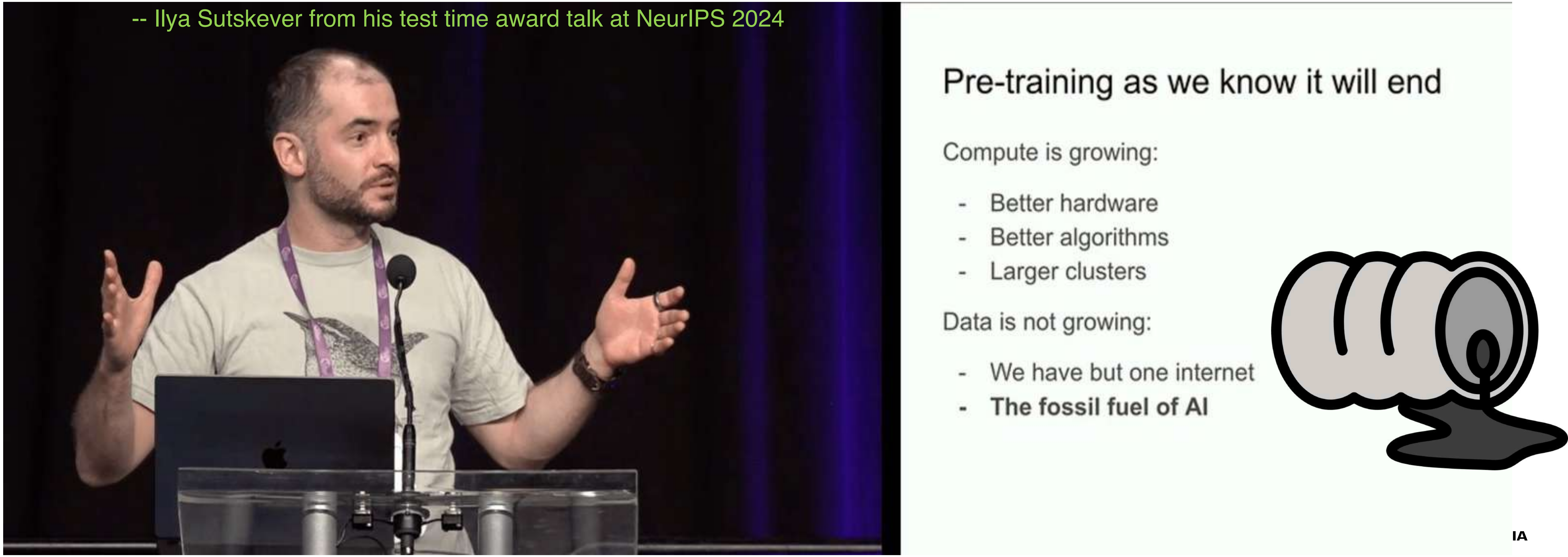
Ali Hatamizadeh[†1], **Syeda Nahida Akter**[†2*], **Shrimai Prabhumoye**[†1,3], **Jan Kautz**[1], **Mostofa Patwary**[1], **Mohammad Shoeybi**[1], **Bryan Catanzaro**[1], **Yejin Choi**[1,4]

NVIDIA[1], Carnegie Mellon University[2], Boston University[3], Stanford University[4]

`ahatamizadeh@nvidia.com`, `sprabhumoye@nvidia.com`

The End of Ever More Pretraining

- Scaling compute is no longer the bottleneck
- High-quality data is finite and increasingly exhausted
- The era shifts from scaling data to extracting more value per token



-- Ilya Sutskever from his test time award talk at NeurIPS 2024

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

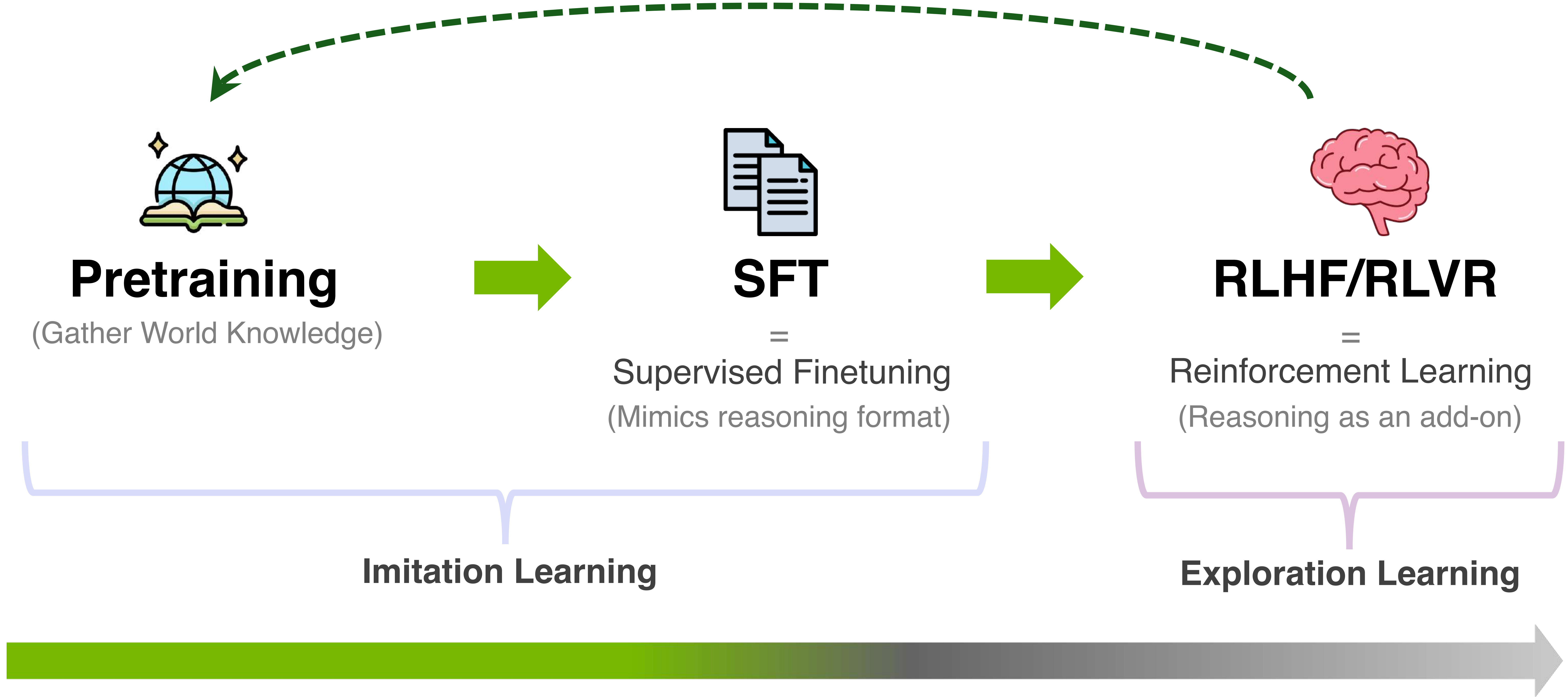
Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

The Problem with Standard LLM Training

Reasoning is an afterthought — we can do better

Q1: Can reasoning be baked in earlier during pretraining — not just added post-hoc?



Pretraining

(Gather World Knowledge)

SFT

=

Supervised Finetuning

(Mimics reasoning format)

RLHF/RLVR

=

Reinforcement Learning

(Reasoning as an add-on)

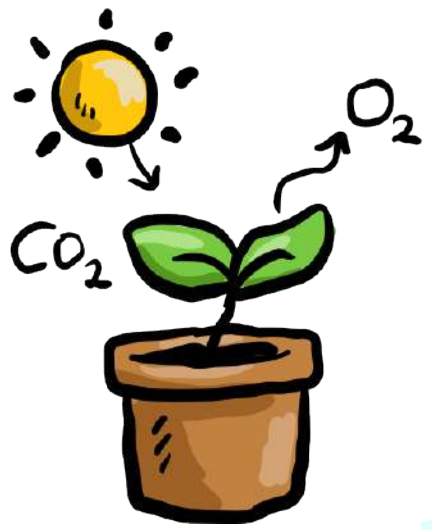
Imitation Learning

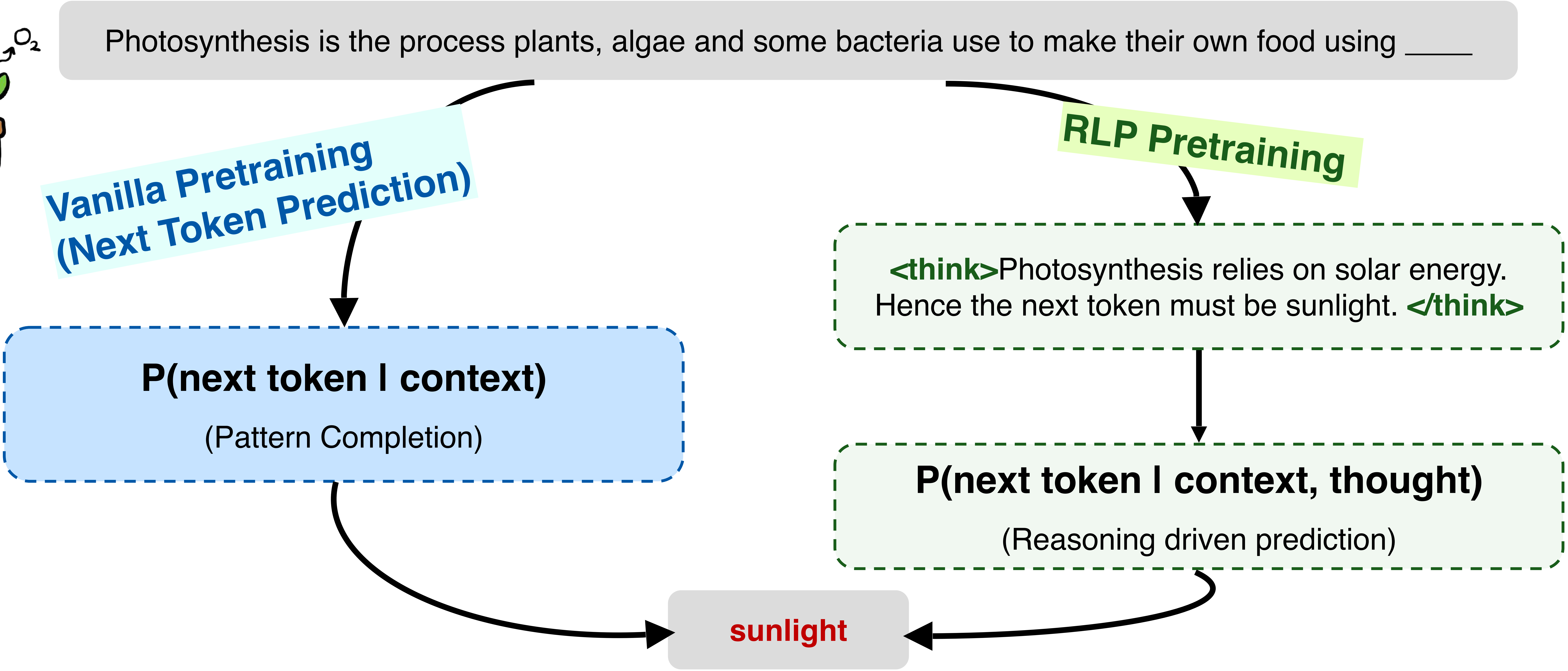
Exploration Learning

Q2: Do gains from early reasoning exposure persist through post-training — or get washed out?

Vanilla Pretraining vs RLP Pretraining

Same context — but RLP induces reasoning.





Photosynthesis is the process plants, algae and some bacteria use to make their own food using ____

Vanilla Pretraining (Next Token Prediction)

P(next token | context)

(Pattern Completion)

RLP Pretraining

<think>Photosynthesis relies on solar energy. Hence the next token must be sunlight. **</think>**

P(next token | context, thought)

(Reasoning driven prediction)

sunlight

Key difference: RLP produces an explicit reasoning trace before predicting the token — making the "why" visible and trainable, not just the final answer.

From NTP to RLP: Teaching Models to Think First

Treat CoT as an action; reward by information gain; update only thought tokens.

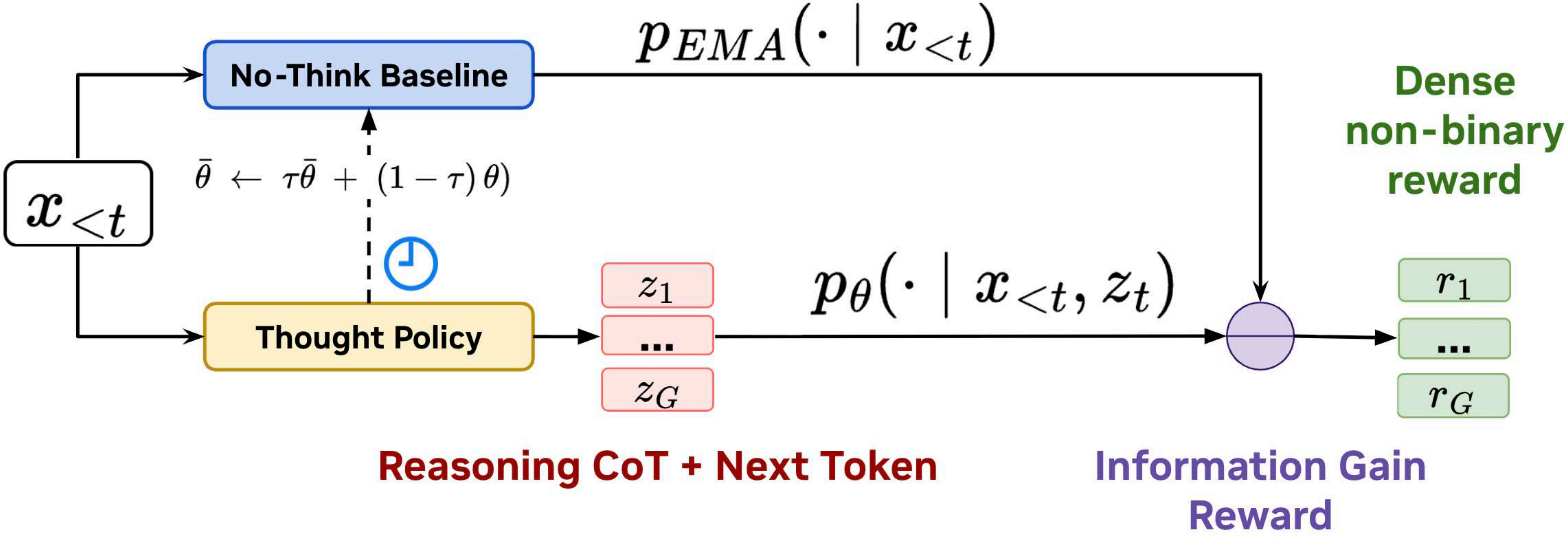
How RLP works

- Sample a thought before predicting the next token.
- Compare likelihoods with vs. without the thought

$$r(c_t) = \log p_\theta\big(x_t \mid x_{<t}, c_t\big) \; - \; \log \bar{p}_\phi\big(x_t \mid x_{<t}\big)$$

- Verifier-free, position-wise credit at every step.

RLP Framework



$x_{<t}$

No-Think Baseline

$p_{EMA}(\cdot \mid x_{<t})$

$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\,\theta)$

Thought Policy

z_1

...

z_G

$p_\theta(\cdot \mid x_{<t}, z_t)$

Dense non-binary reward

r_1

...

r_G

Reasoning CoT + Next Token

Information Gain Reward

Training Recipe

- **Single network** acts as policy + reasoned predictor.
- **EMA "no-think" teacher** with slow lag $\phi \leftarrow \tau\phi + (1-\tau)\,\theta, \qquad \tau = 0.999$
- **Group-relative advantages** with inclusive-mean correction

$$\bar{r} = \frac{1}{G}\sum_{j=1}^{G} r\!\left(c_t^{(j)}\right), \qquad A^{(i)} = \frac{G}{G-1}\left(r\!\left(c_t^{(i)}\right) - \bar{r}\right)$$

- **Clipped, per-token surrogate** (GRPO-style) on **thought tokens** only

$$\mathscr{L}_{\text{clip}} = -\,\mathbb{E}\left[\frac{1}{|\,c_t^{(i)}\,|}\sum_{u}\min\!\left(\rho_u^{(i)}A^{(i)}, \mathrm{clip}\!\left(\rho_u^{(i)}; 1-\epsilon_\ell, 1+\epsilon_h\right)A^{(i)}\right)\right]$$

with per-token importance ratios

$$\rho_u^{(i)} = \exp\!\left(\log \pi_\theta\!\left(\ell_u^{(i)} \mid x_{<t}, \ell_{1:u-1}^{(i)}\right) - \log \pi_{\theta_{\text{old}}}\!\left(\ell_u^{(i)} \mid x_{<t}, \ell_{1:u-1}^{(i)}\right)\right)$$

Leverage Early Reasoning with “RLP"

Q1

Can RLP improve the reasoning ability of a base model without any prior task-specific tuning? (And better than Base + CPT?)
Does the gain of RLP survive even after SFT+RLVR?

Q2

Can RLP, when applied to an earlier pre-training checkpoint with 200B less tokens (to match the pretraining FLOPs), match the performance of a fully trained base model?

Training Ingredien

Model: Qwen3-1.7B-Base
Checkpoint: Final model
Token Budget: 1B Tokens
CPT Data ($\mathscr{D}$):

- General Pretraining Corpora (
Pretraining-Dataset)

Comparison with:

- Base model
  - **BASE**
  - **BASE+POST**
- **Token Matched:** NTP-based CPT on base model
  - **BASE+CPT**
  - **BASE+CPT+POST**

Training Ingredients

Model: NVIDIA-Nemotron-Nano-12B-v2-Bas
Checkpoint: Model trained till 19.8T tokens
Token Budget: 250M Tokens
CPT Data ($\mathscr{D}$):

- General Pretraining Corpora (Nemotron-Pretraining-Dataset)

Comparison with:

- FLOP Matched: Base model (20T)
  - **BASE**
  - **BASE+POST**

RLP with *200B* less tokens!

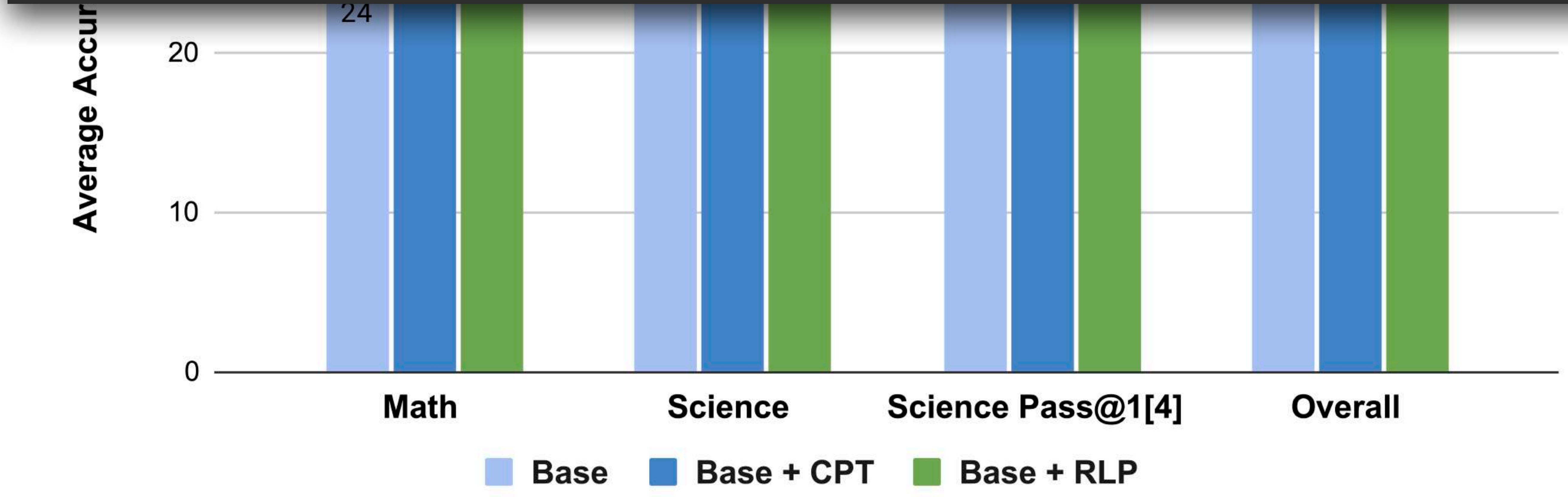
RLP shows a significant improvement on Qwen3-1.7B-Base

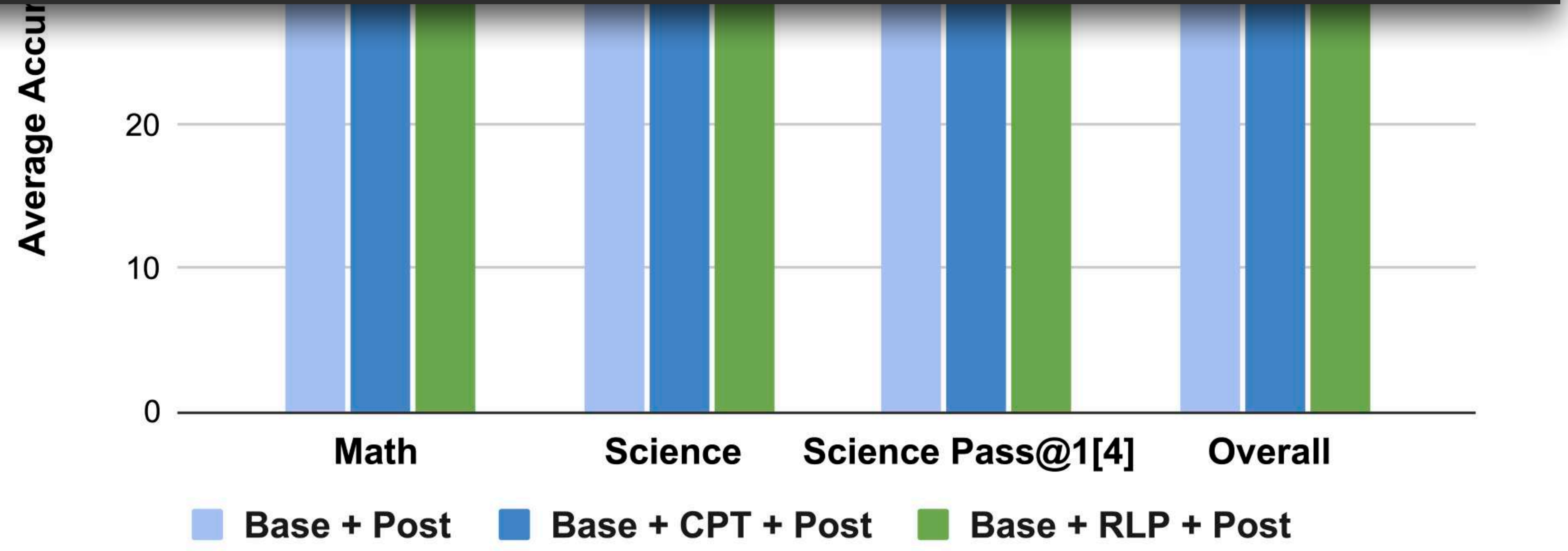
RLP outperforms BASE by *+19%* and **CPT by *+17%*** on average across math and science benchmarks.

After identical SFT + RLVR, **RLP compounds its advantage: *+8%*** relative over BASE+Post; ***+7%*** relative over CPT+Post

Unlike NPT or continuous pretraining, RLP instills reasoning during pretraining itself

The early advantages compound through post training, giving models stronger and more robust reasoning capabilities.





RLP shows a significant improvement on Qwen3-1.7B-Base

Does the improvement sustain under compute equivalent baselines?

Even with flop matching training, **RLP outperforms CPT** (exposed to **35× more data**) by ***+14%*** on average!

$$T_{flop} = (n \times l_{seq} \times bs \times iters) + T_{inp}$$

n = rollouts

bs = batch size

l_{seq} = sequence length

Training Ingredients

Model: Qwen3-1.7B-Base

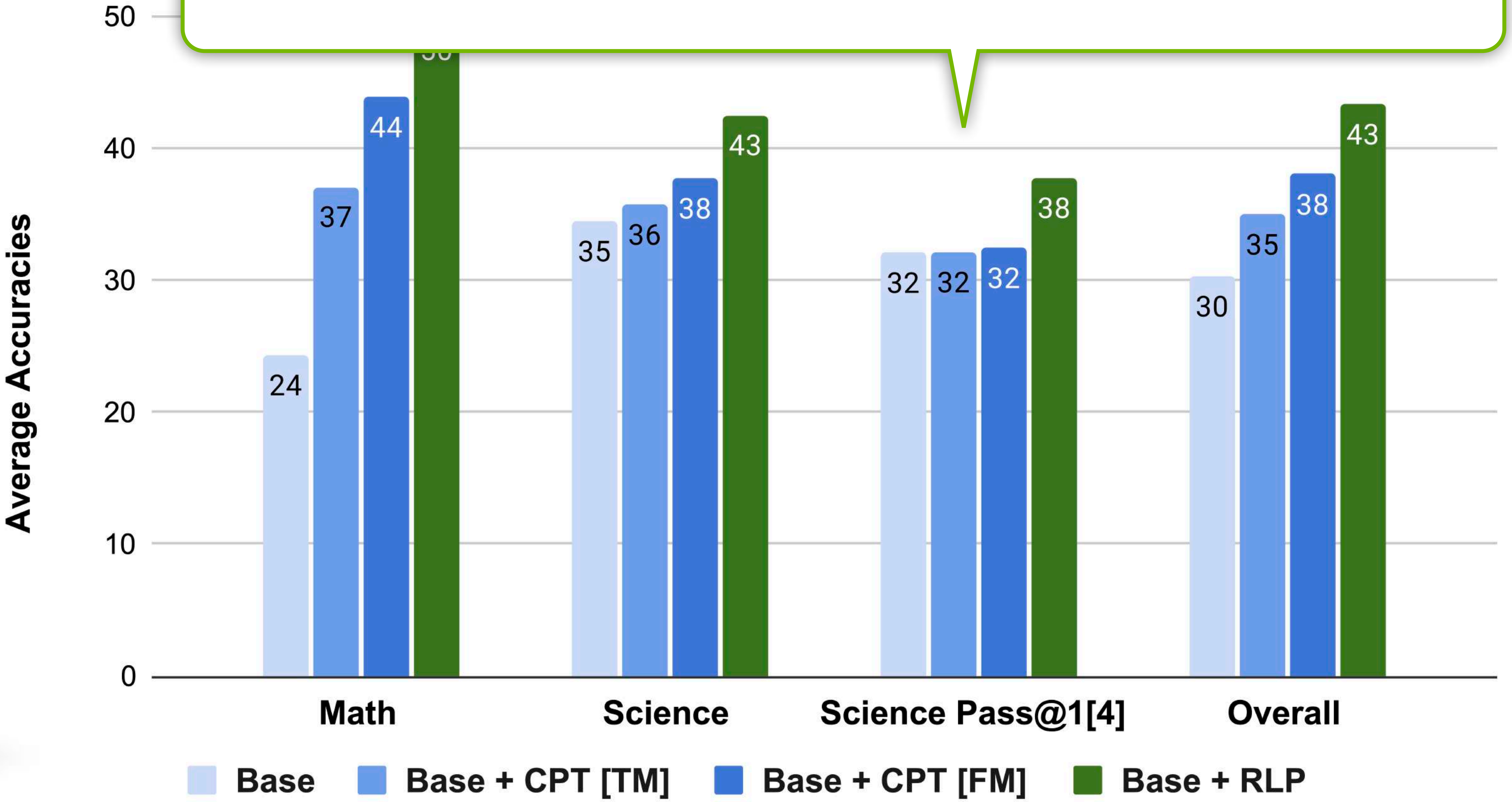
Checkpoint: Final model

Token Matching (T_{inp}): 170M Tokens

Flop Matching (T_{flop}): 6B Tokens

CPT Data ($\mathscr{D}$):

- Nemotron-CrossThink



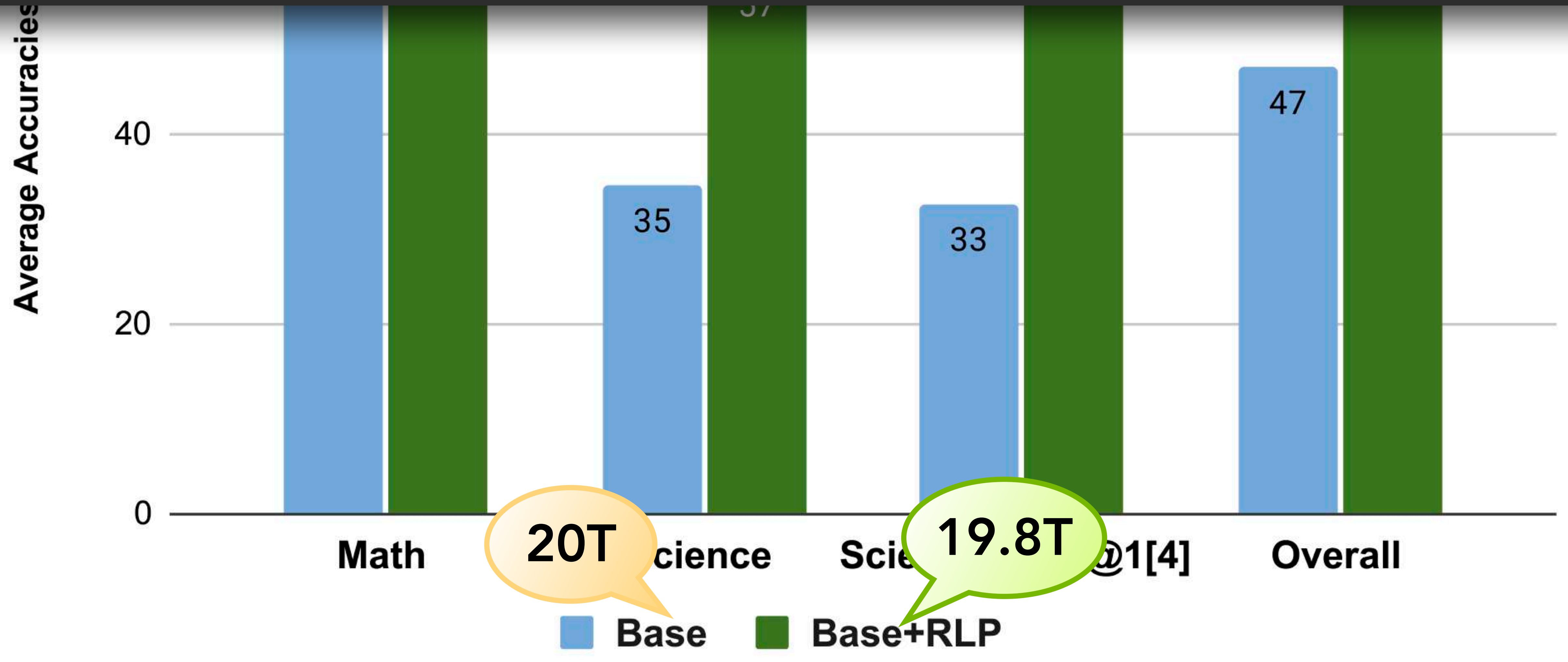
RLP scales with LLM size (12B) and frameworks

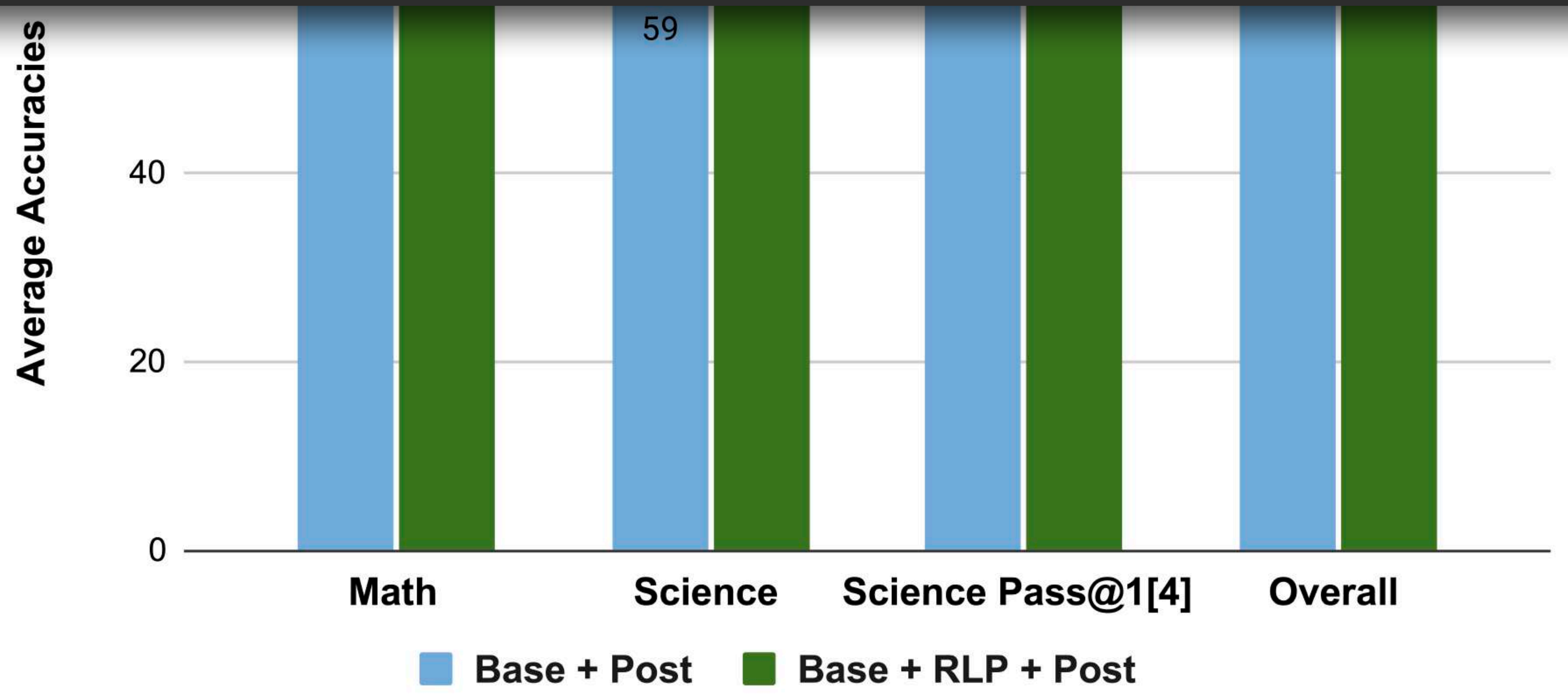
RLP, trained on **~200B fewer** tokens, achieves a **35% average gain**, with the largest boost in science (**+23% absolute**), showing robust cross-domain benefits at

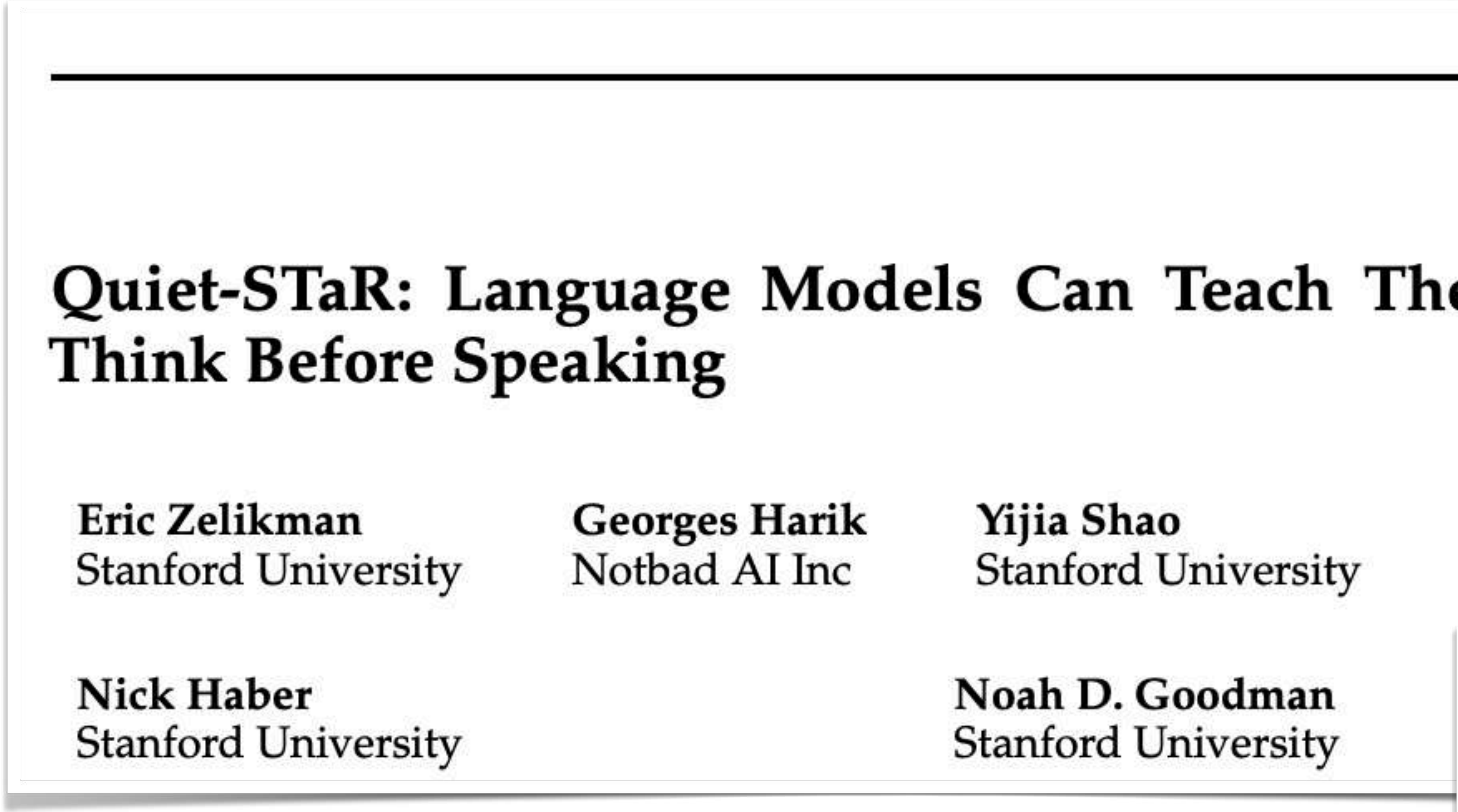
After identical post-training, **RLP** outperforms BASE by ***+2% absolute margin,*** which has seen **~200B** more tokens during pretraining.

The benefits of RLP persist and even amplify when scaling to larger models and across architectures.

Efficiency meets intelligence—proof that reasoning, not sheer scale, drives progress.







Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking

Eric Zelikman
Stanford University

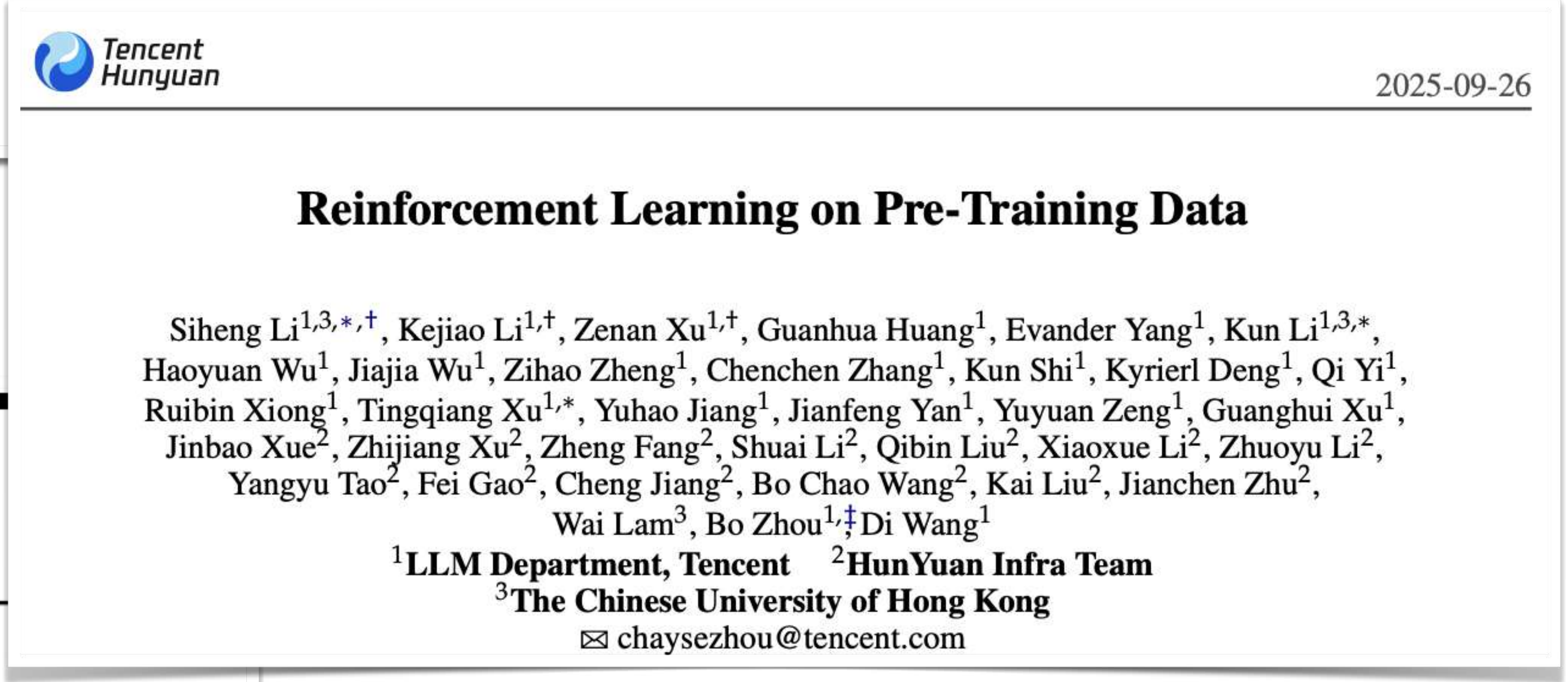
Georges Harik
Notbad AI Inc

Yijia Shao
Stanford University

Varuna Jayasiri
Notbad AI Inc

Nick Haber
Stanford University

Noah D. Goodman
Stanford University



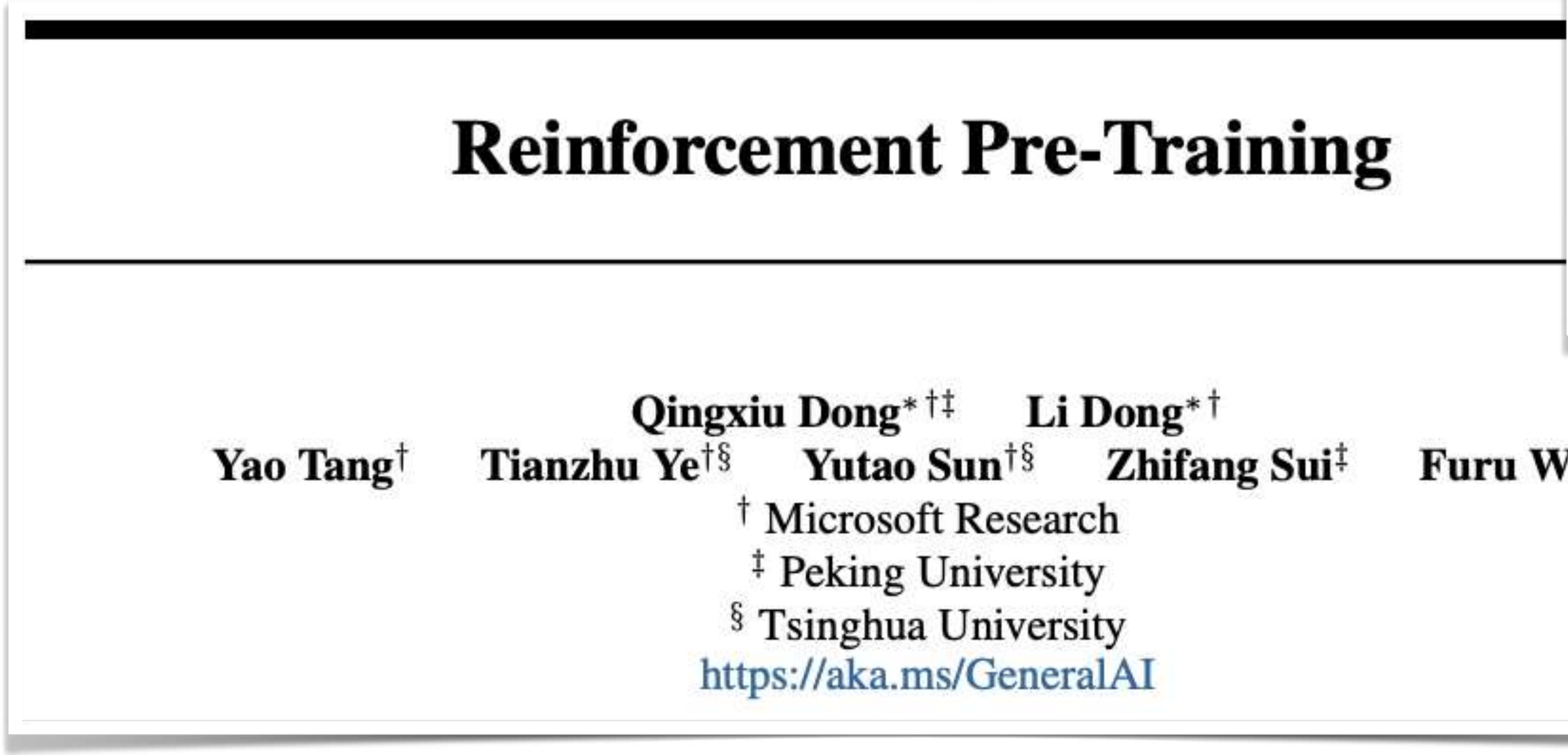
Tencent Hunyuan

2025-09-26

Reinforcement Learning on Pre-Training Data

Siheng Li[1,3,*,†], Kejiao Li[1,†], Zenan Xu[1,†], Guanhua Huang[1], Evander Yang[1], Kun Li[1,3,*], Haoyuan Wu[1], Jiajia Wu[1], Zihao Zheng[1], Chenchen Zhang[1], Kun Shi[1], Kyrierl Deng[1], Qi Yi[1], Ruibin Xiong[1], Tingqiang Xu[1,*], Yuhao Jiang[1], Jianfeng Yan[1], Yuyuan Zeng[1], Guanghui Xu[1], Jinbao Xue[2], Zhijiang Xu[2], Zheng Fang[2], Shuai Li[2], Qibin Liu[2], Xiaoxue Li[2], Zhuoyu Li[2], Yangyu Tao[2], Fei Gao[2], Cheng Jiang[2], Bo Chao Wang[2], Kai Liu[2], Jianchen Zhu[2], Wai Lam[3], Bo Zhou[1,‡] Di Wang[1]

[1]LLM Department, Tencent **[2]HunYuan Infra Team**
[3]The Chinese University of Hong Kong
✉ chaysezhou@tencent.com



Reinforcement Pre-Training

Qingxiu Dong[*,†,‡] **Li Dong**[*,†]
Yao Tang[†] **Tianzhu Ye**[†,§] **Yutao Sun**[†,§] **Zhifang Sui**[‡] **Furu Wei**[†,◇]
[†] Microsoft Research
[‡] Peking University
[§] Tsinghua University
https://aka.ms/GeneralAI

RLP: The Core Idea

Chain-of-thought as an exploratory action during pretraining

```
Reward(thought) = log P(next token | context + thought) − log P(next token | context alone)
```

💡

Verifier-Free

No external judge needed. The reward signal comes directly from information gain on the next token — applicable to any ordinary pretraining text.

⚡

Dense Reward

Position-wise credit assigned at every token where reasoning helps. Unlike sparse post-training rewards, RLP provides continuous feedback throughout training.

📈

Scalable

Works on the same massive pretraining corpora used for standard training. No specialized datasets required — reasoning emerges from plain text.

RLP bridges next-token prediction and the emergence of useful chain-of-thought reasoning — all during pretraining.

Front-Loading Reasoning

The Synergy between Pretraining and Post-Training Data

— ICLR 2026 —

Syeda Nahida Akter[2*], **Shrimai Prabhumoye**[1,3], **Eric Nyberg**[2], **Mostofa Patwary**[1],
Mohammad Shoeybi[1], **Yejin Choi**[1,4], **Bryan Catanzaro**[1]
NVIDIA[1], Carnegie Mellon University[2], Boston University[3], Stanford University[4]
sakter@andrew.cmu.edu, sprabhumoye@nvidia.com

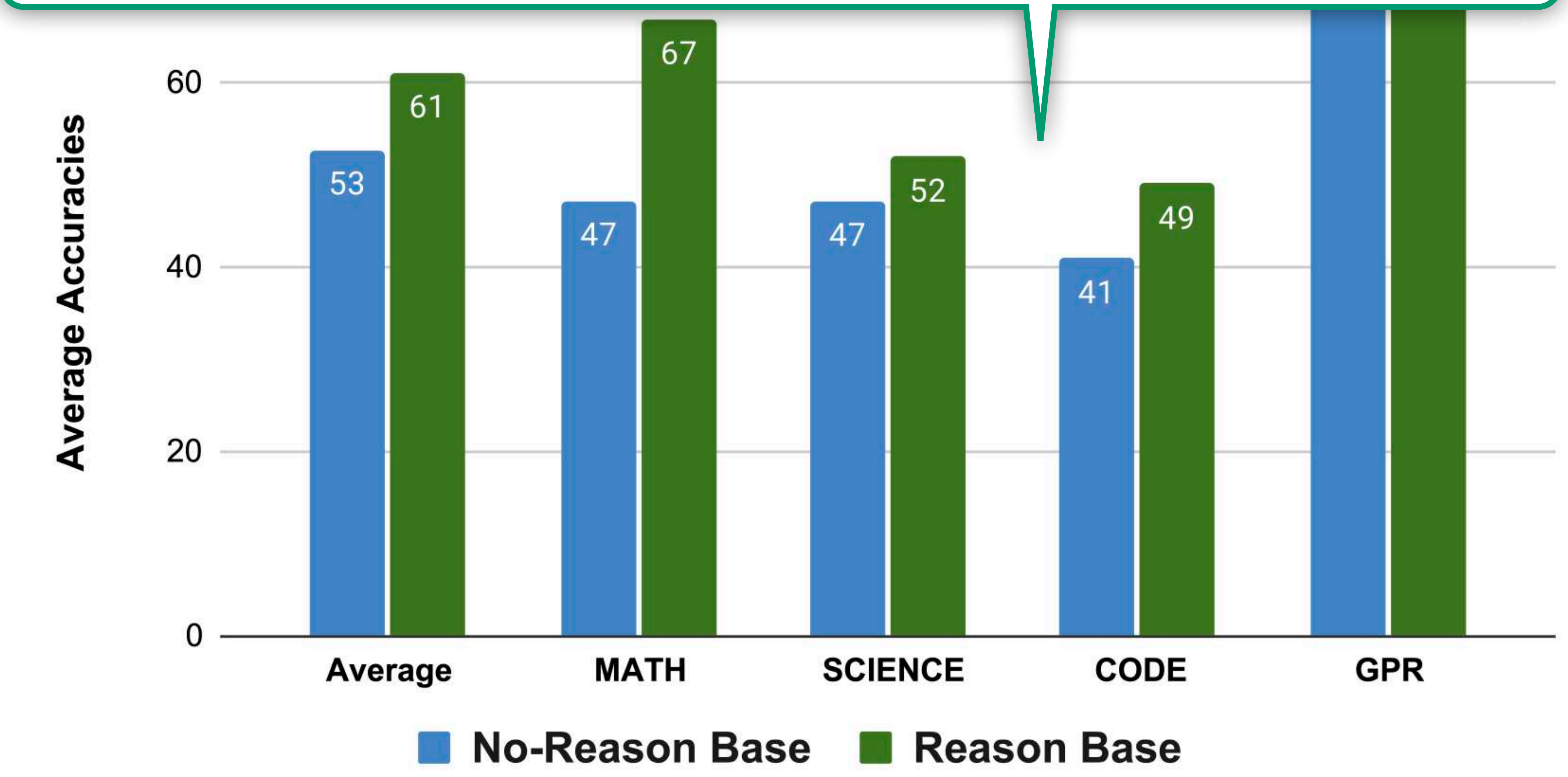
Front Loading Reasoning

FLR systematically injects reasoning-style data at different phases of training—pretraining, SFT, RL —while varying its diversity, quantity, and quality.

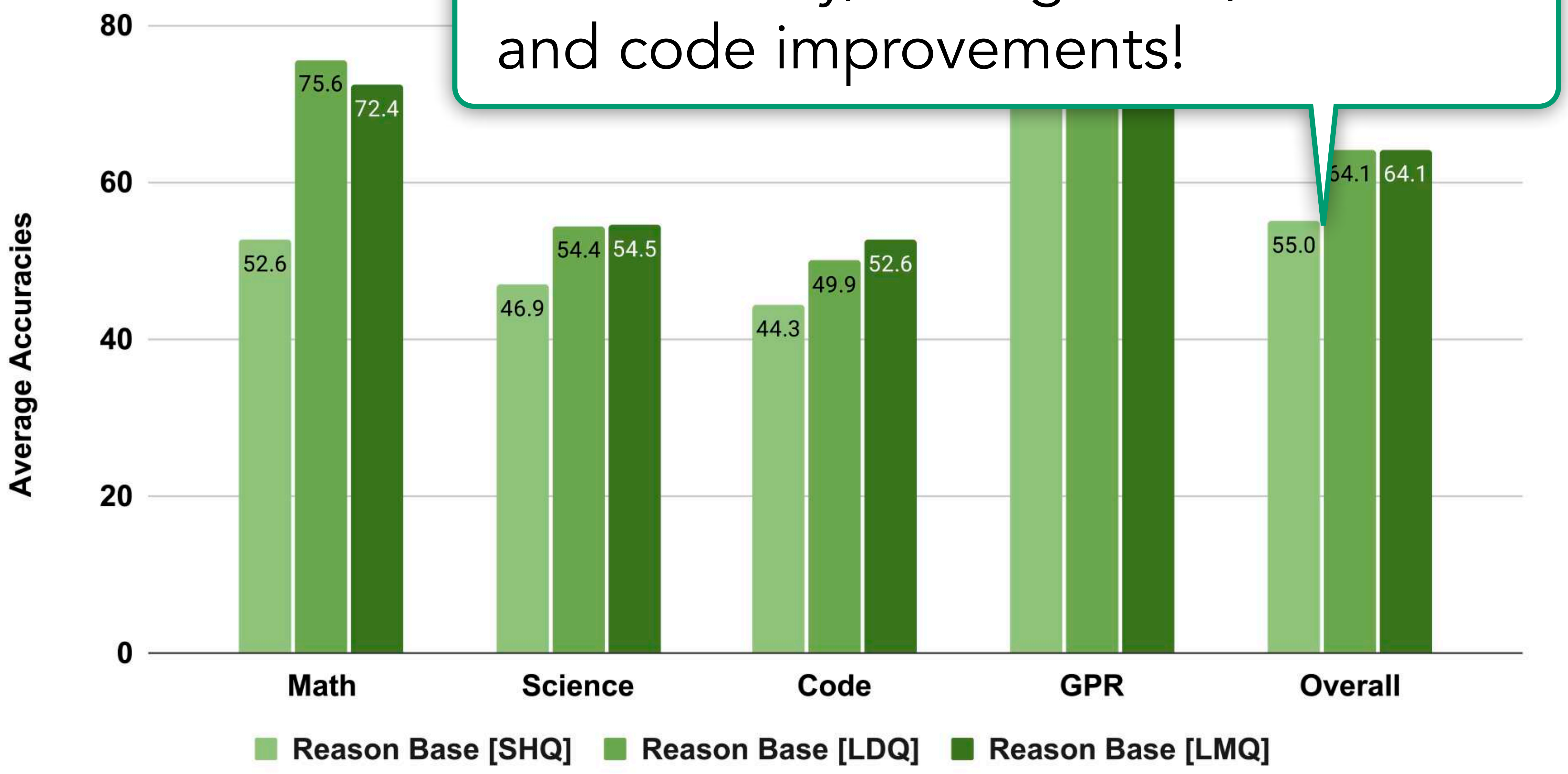
FLR

RQ1: Is the inclusion of reasoning data in pretraining beneficial for the base model?
Lesson I: Yes, **highly beneficial**, and the benefit scales with **diversity and quantity**!

Using reasoning data in pretraining yields an absolute ***+16% average improvement*** over the no-reason baseline.



Pretraining gains improve with scale and diversity, driving math, science and code improvements!

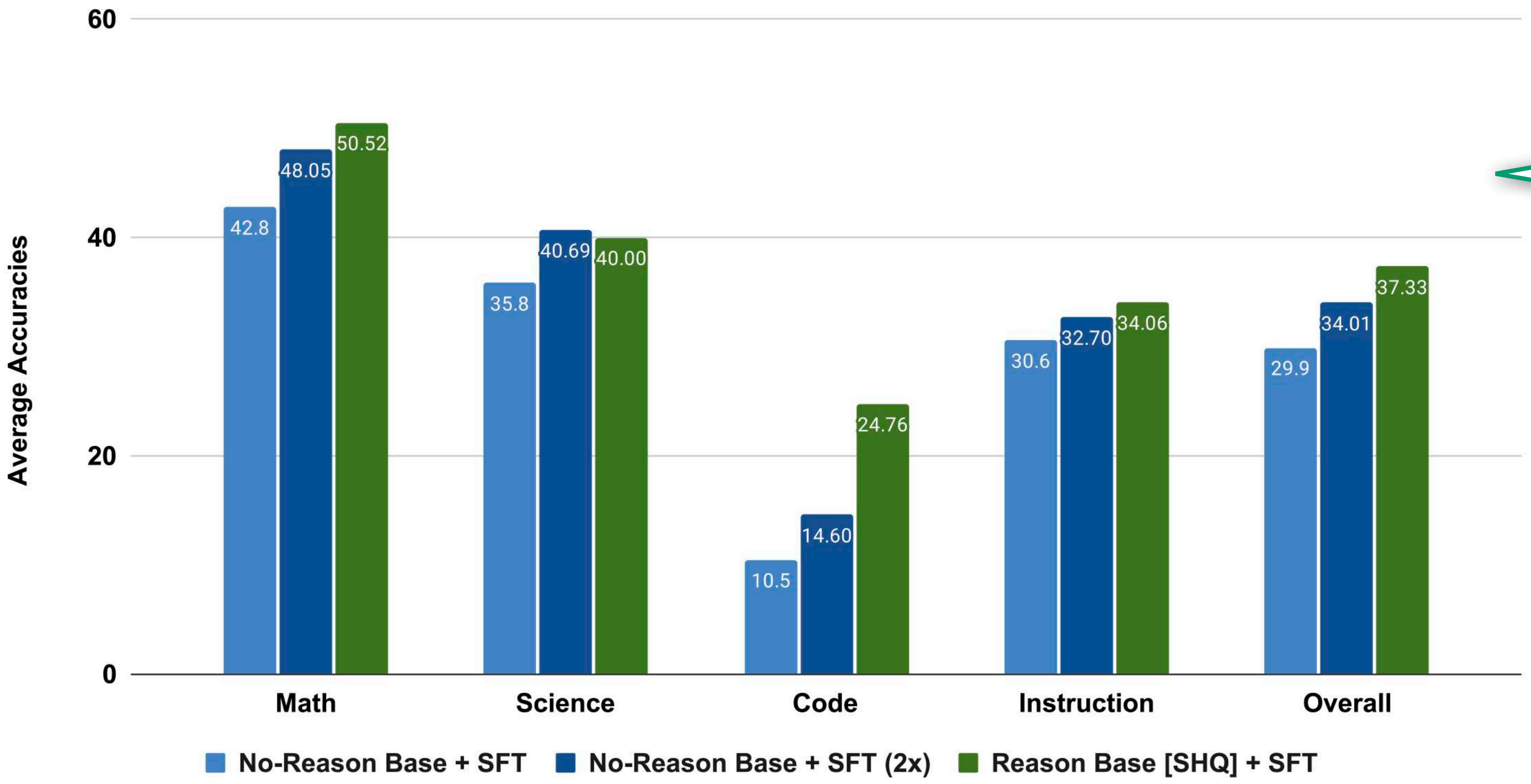


Front Loading Reasoning

FLR systematically injects reasoning-style data at different phases of training—pretraining, SFT, RL—while varying its diversity, quantity, and quality.

FLR

RQ3: Can a model with no-reason pretraining "catch up" by getting more SFT compute?
Lesson IV: **No**! Early reasoning builds an **irreplaceable** foundation that SFT can't replicate



Doubling SFT data for the baseline improves performance by **+4.09%**, but it still falls short of even the weakest reasoning-pretrained model (**+3.32%**).

Front Loading Reasoning

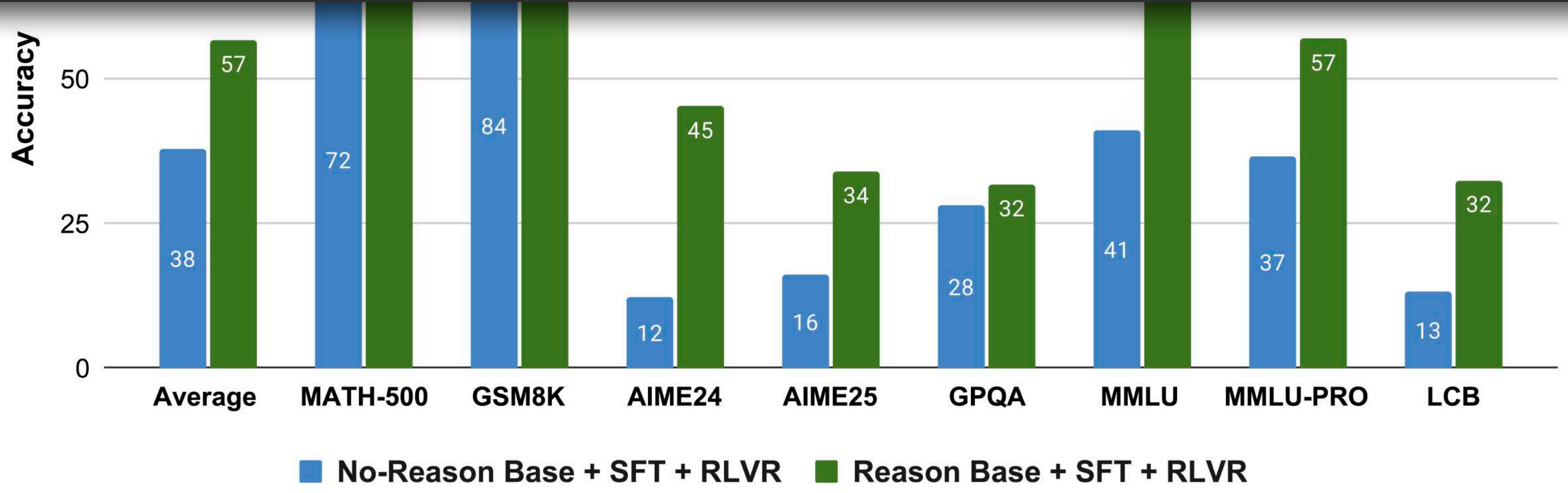
FLR systematically injects reasoning-style data at different phases of training—pretraining, SFT, RL —while varying its diversity, quantity, and quality.

FLR

RQ5: Does the gain matter after heavy RLVR?

Lesson VI: **Yes**! "Front-loading" reasoning data creates a **durable, compounding gain**

- Front-loading reasoning data into pretraining creates a durable, compounding advantage.
- High-quality pretraining data can have a latent effect unlocked by SFT
- The optimal data strategy is asymmetric: prioritize diversity in pretraining and quality in SFT.



Reason-Base model finished with a stunning **+19%** average lead. On the most difficult competition math problems (AIME), this advantage ballooned to a **+39.3%** gain.

Concluding Remarks

David vs. Goliath: the Art of Scaling Intelligence in the Era of Extreme-Scale Neural Models

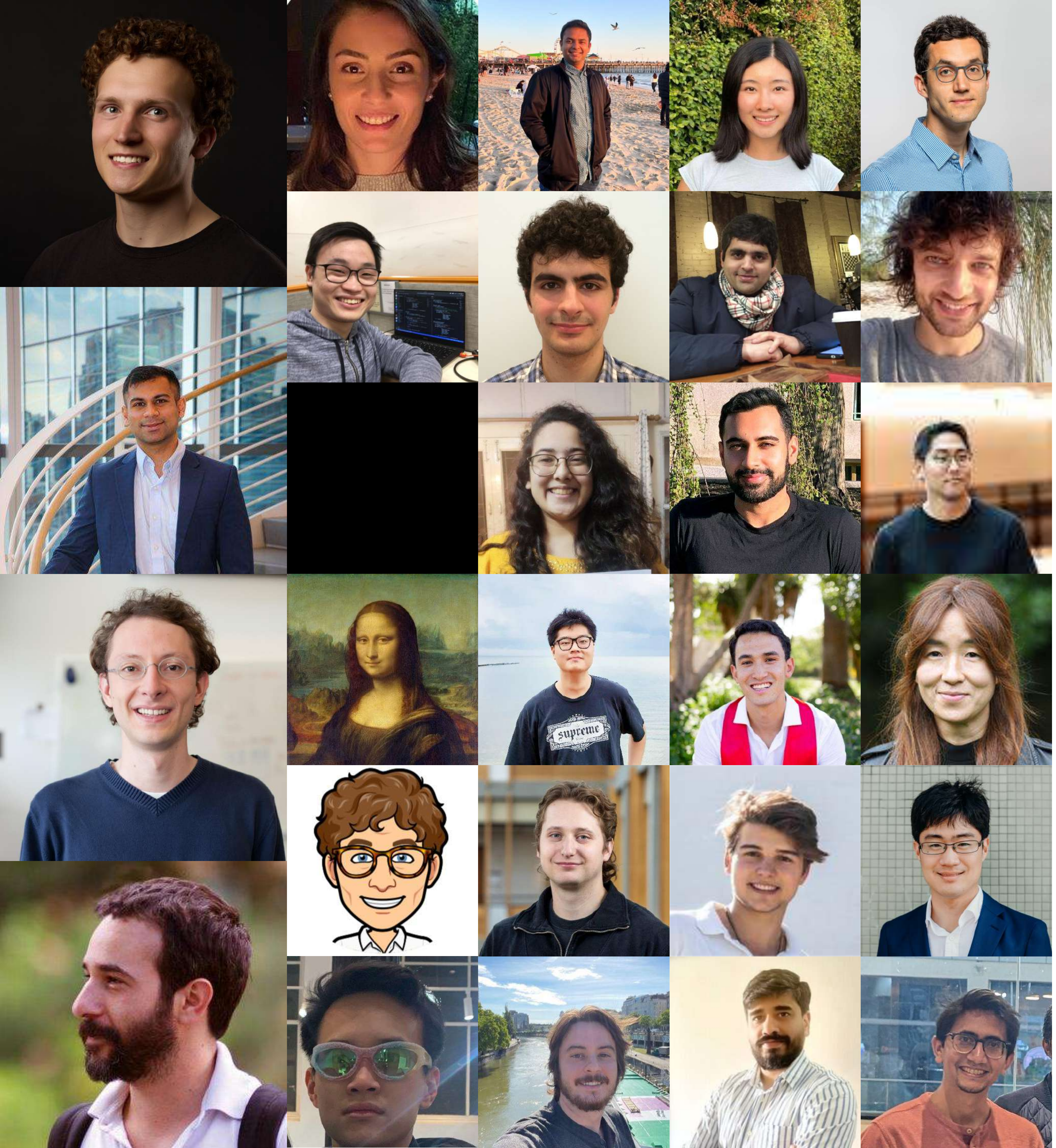
Three components to innovate:

Unconventional data 🔥

Unconventional algorithms 🚀

Unconventional collaboration 🌏

- implicit: open-science & open-source community
- explicit: cross-institutional, cross-boarder collaborations

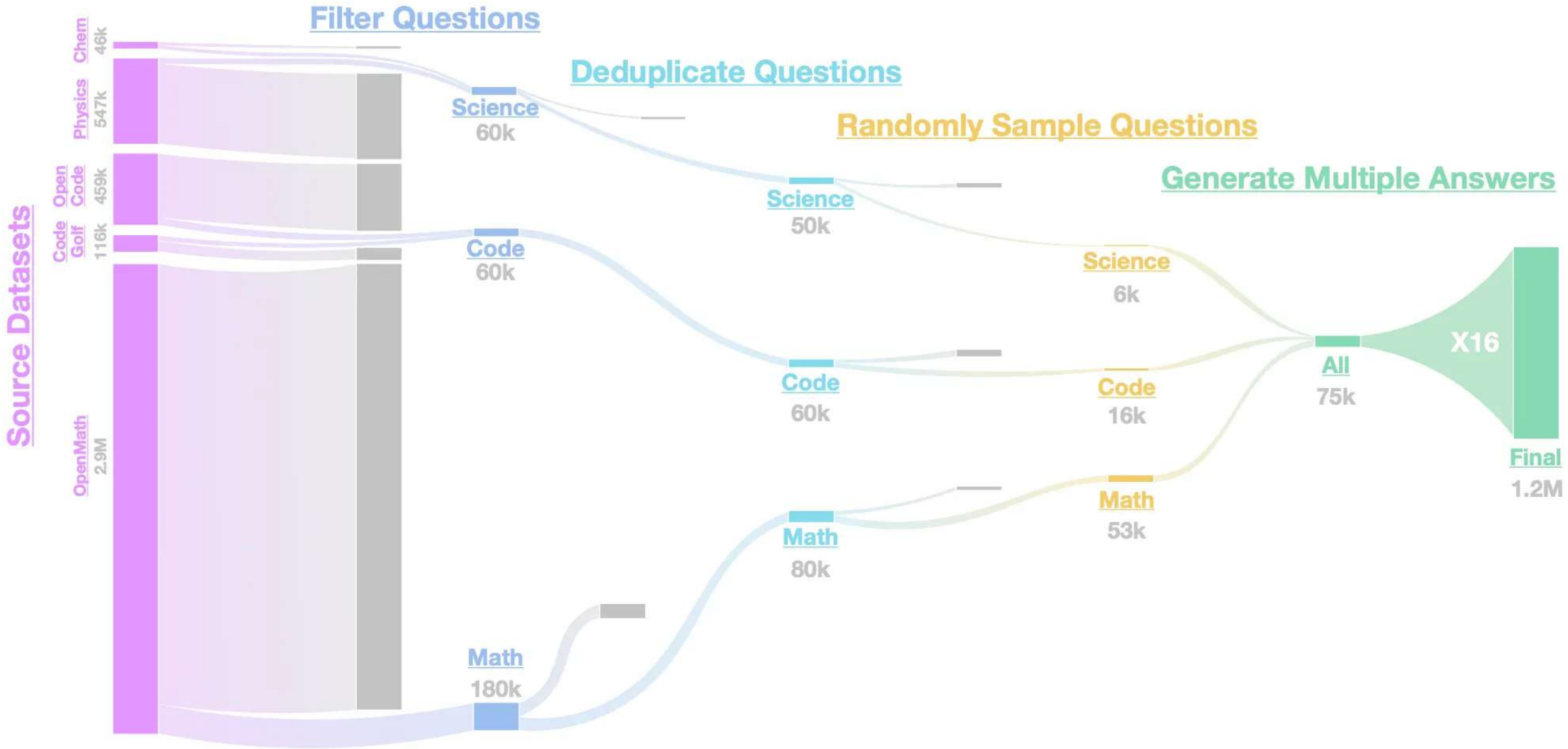


The OpenThoughts Team

Etash Guha*, Ryan Marten*, Sedrick Keh*, Negin Raoof*, Georgios Smyrnis*, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, Ludwig Schmidt*

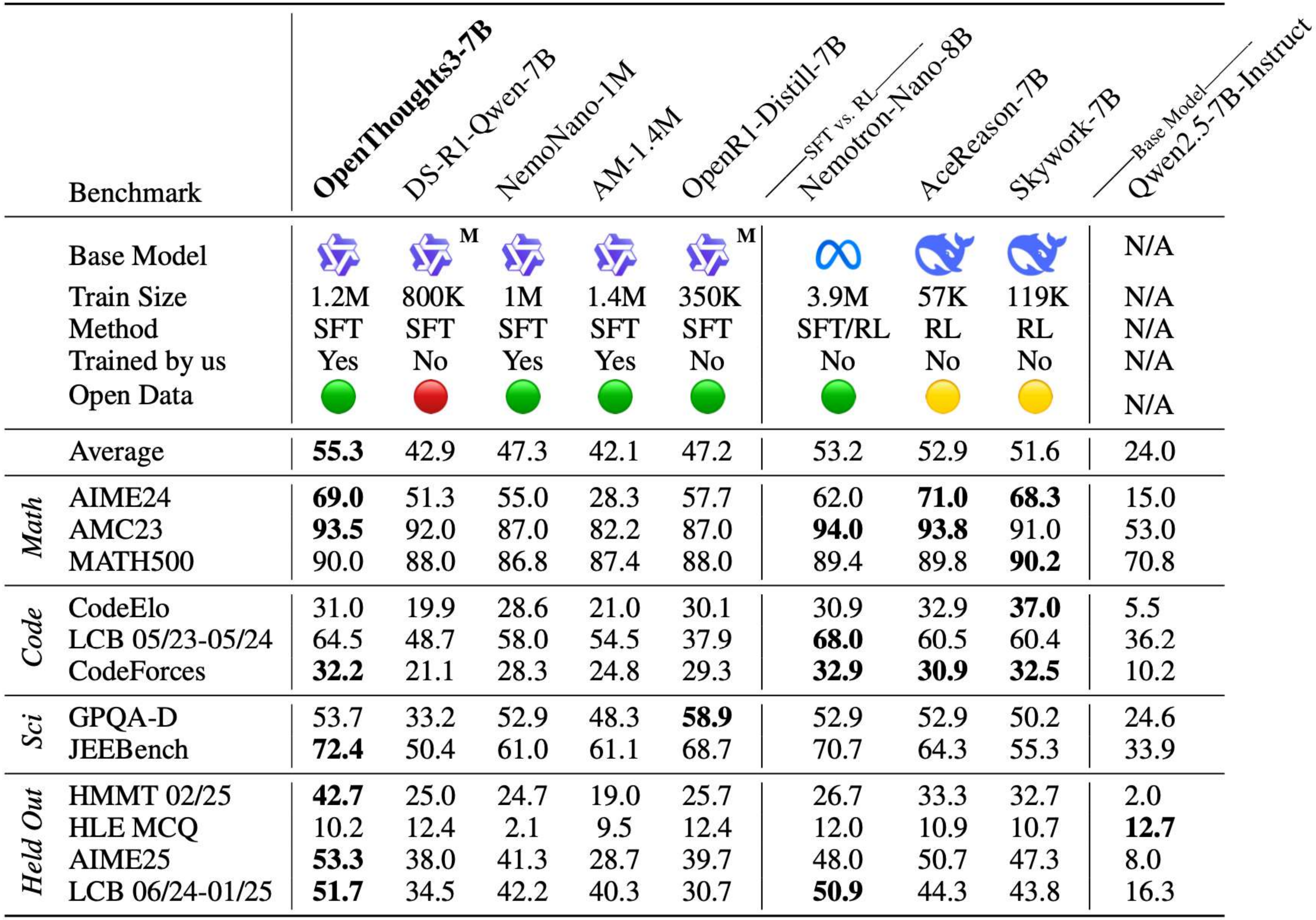
And several others with ongoing contributions

OpenThoughts3-1M



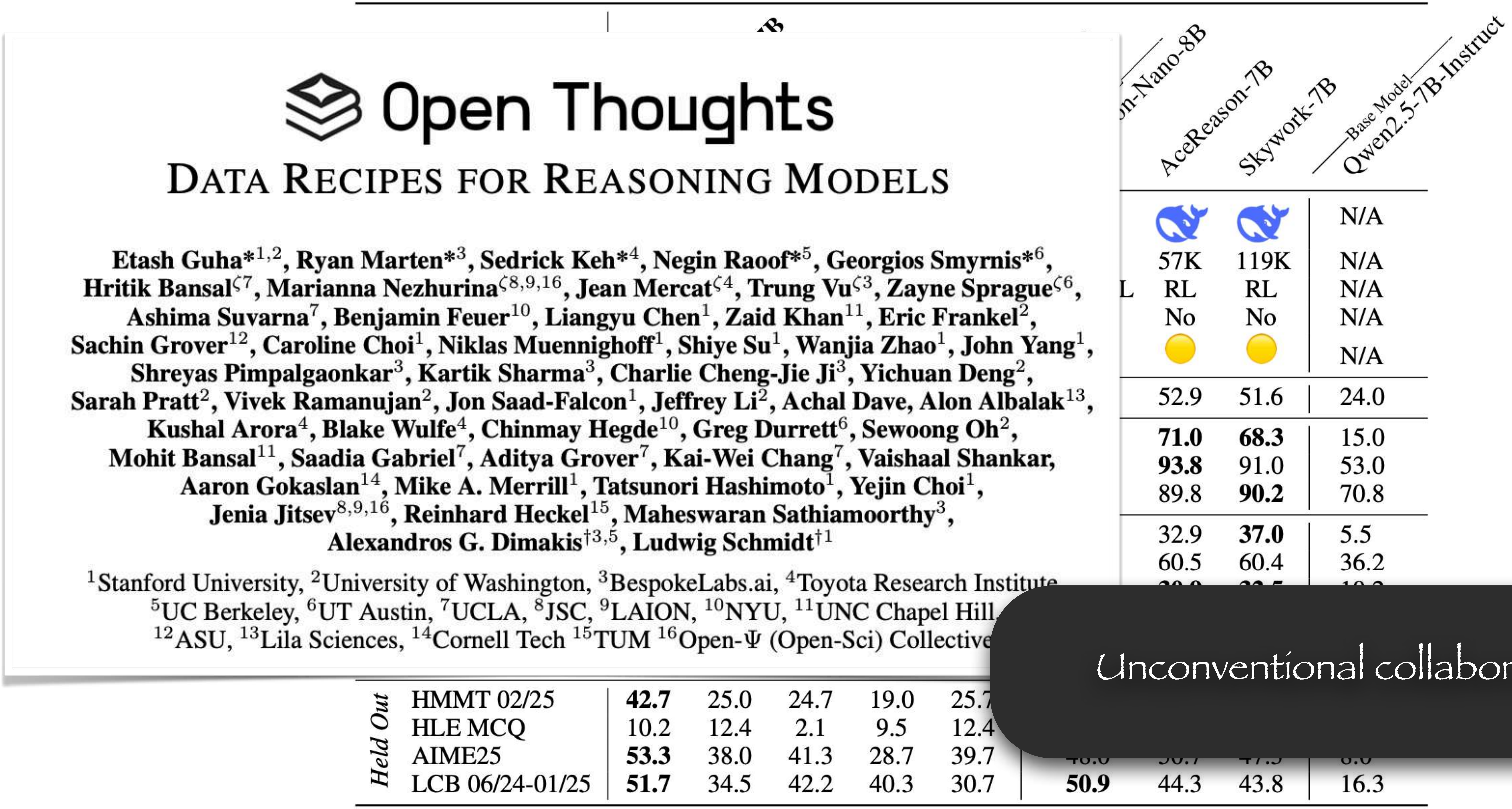
OpenThoughts3 is the SOTA reasoning dataset recipe

AIME 2025
LiveCodeBench
GPQA Diamond
Accuracy (%)
0
10
20
30
40
50
60
1K
10K
100K
1M
Dataset Size
OpenThoughts3
Nemotron Nano
AM
s1.1
LIMO
Qwen-2.5-7B-Instruct



							SFT vs. RL			Base Model
	Benchmark	**OpenThoughts3-7B**	DS-R1-Qwen-7B	NemoNano-1M	AM-1.4M	OpenR1-Distill-7B	Nemotron-Nano-8B	AceReason-7B	Skywork-7B	Qwen2.5-7B-Instruct
	Base Model		M			M				N/A
	Train Size	1.2M	800K	1M	1.4M	350K	3.9M	57K	119K	N/A
	Method	SFT	SFT	SFT	SFT	SFT	SFT/RL	RL	RL	N/A
	Trained by us	Yes	No	Yes	Yes	No	No	No	No	N/A
	Open Data	🟢	🔴	🟢	🟢	🟢	🟢	🟡	🟡	N/A
	Average	**55.3**	42.9	47.3	42.1	47.2	53.2	52.9	51.6	24.0
Math	AIME24	**69.0**	51.3	55.0	28.3	57.7	62.0	**71.0**	**68.3**	15.0
	AMC23	**93.5**	92.0	87.0	82.2	87.0	**94.0**	**93.8**	91.0	53.0
	MATH500	90.0	88.0	86.8	87.4	88.0	89.4	89.8	**90.2**	70.8
Code	CodeElo	31.0	19.9	28.6	21.0	30.1	30.9	32.9	**37.0**	5.5
	LCB 05/23-05/24	64.5	48.7	58.0	54.5	37.9	**68.0**	60.5	60.4	36.2
	CodeForces	**32.2**	21.1	28.3	24.8	29.3	**32.9**	**30.9**	**32.5**	10.2
Sci	GPQA-D	53.7	33.2	52.9	48.3	**58.9**	52.9	52.9	50.2	24.6
	JEEBench	**72.4**	50.4	61.0	61.1	68.7	70.7	64.3	55.3	33.9
Held Out	HMMT 02/25	**42.7**	25.0	24.7	19.0	25.7	26.7	33.3	32.7	2.0
	HLE MCQ	10.2	12.4	2.1	9.5	12.4	12.0	10.9	10.7	**12.7**
	AIME25	**53.3**	38.0	41.3	28.7	39.7	48.0	50.7	47.3	8.0
	LCB 06/24-01/25	**51.7**	34.5	42.2	40.3	30.7	**50.9**	44.3	43.8	16.3

Super effortful SFT can win over effortful RL



Open Thoughts

Data Recipes for Reasoning Models

Etash Guha*[1,2], Ryan Marten*[3], Sedrick Keh*[4], Negin Raoof*[5], Georgios Smyrnis*[6], Hritik Bansal[ζ7], Marianna Nezhurina[ζ8,9,16], Jean Mercat[ζ4], Trung Vu[ζ3], Zayne Sprague[ζ6], Ashima Suvarna[7], Benjamin Feuer[10], Liangyu Chen[1], Zaid Khan[11], Eric Frankel[2], Sachin Grover[12], Caroline Choi[1], Niklas Muennighoff[1], Shiye Su[1], Wanjia Zhao[1], John Yang[1], Shreyas Pimpalgaonkar[3], Kartik Sharma[3], Charlie Cheng-Jie Ji[3], Yichuan Deng[2], Sarah Pratt[2], Vivek Ramanujan[2], Jon Saad-Falcon[1], Jeffrey Li[2], Achal Dave, Alon Albalak[13], Kushal Arora[4], Blake Wulfe[4], Chinmay Hegde[10], Greg Durrett[6], Sewoong Oh[2], Mohit Bansal[11], Saadia Gabriel[7], Aditya Grover[7], Kai-Wei Chang[7], Vaishaal Shankar, Aaron Gokaslan[14], Mike A. Merrill[1], Tatsunori Hashimoto[1], Yejin Choi[1], Jenia Jitsev[8,9,16], Reinhard Heckel[15], Maheswaran Sathiamoorthy[3], Alexandros G. Dimakis[†3,5], Ludwig Schmidt[†1]

[1]Stanford University, [2]University of Washington, [3]BespokeLabs.ai, [4]Toyota Research Institute
[5]UC Berkeley, [6]UT Austin, [7]UCLA, [8]JSC, [9]LAION, [10]NYU, [11]UNC Chapel Hill
[12]ASU, [13]Lila Sciences, [14]Cornell Tech [15]TUM [16]Open-Ψ (Open-Sci) Collective

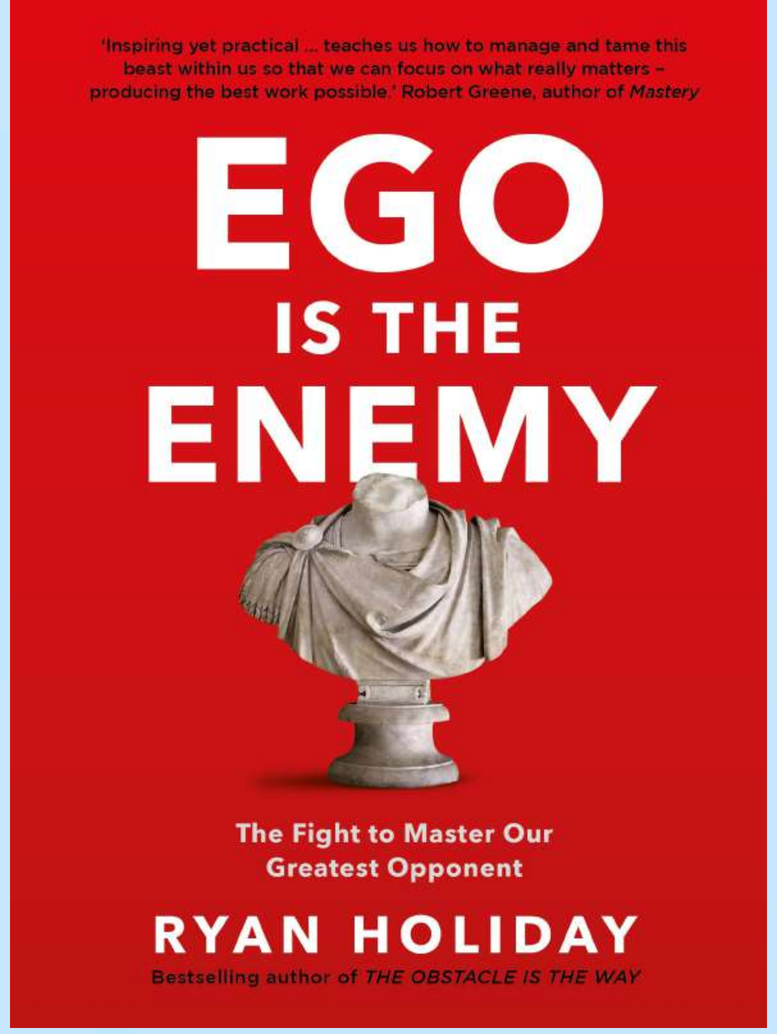
…n-Nano-8B	AceReason-7B	Skywork-7B	Base Model Qwen2.5-7B-Instruct
	🐋	🐋	N/A
	57K	119K	N/A
L	RL	RL	N/A
	No	No	N/A
	●	●	N/A
	52.9	51.6	24.0
	71.0	**68.3**	15.0
	93.8	91.0	53.0
	89.8	**90.2**	70.8
	32.9	**37.0**	5.5
	60.5	60.4	36.2

Held Out	HMMT 02/25	**42.7**	25.0	24.7	19.0	25.7			
	HLE MCQ	10.2	12.4	2.1	9.5	12.4			
	AIME25	**53.3**	38.0	41.3	28.7	39.7			
	LCB 06/24-01/25	**51.7**	34.5	42.2	40.3	30.7	**50.9**	44.3	43.8	16.3

Unconventional collaboration wins!

Super effortful SFT can win over effortful RL

Concluding Thoughts


'Inspiring yet practical … teaches us how to manage and tame this beast within us so that we can focus on what really matters – producing the best work possible.' Robert Greene, author of *Mastery*
EGO
IS THE
ENEMY
The Fight to Master Our Greatest Opponent
RYAN HOLIDAY
Bestselling author of *THE OBSTACLE IS THE WAY*

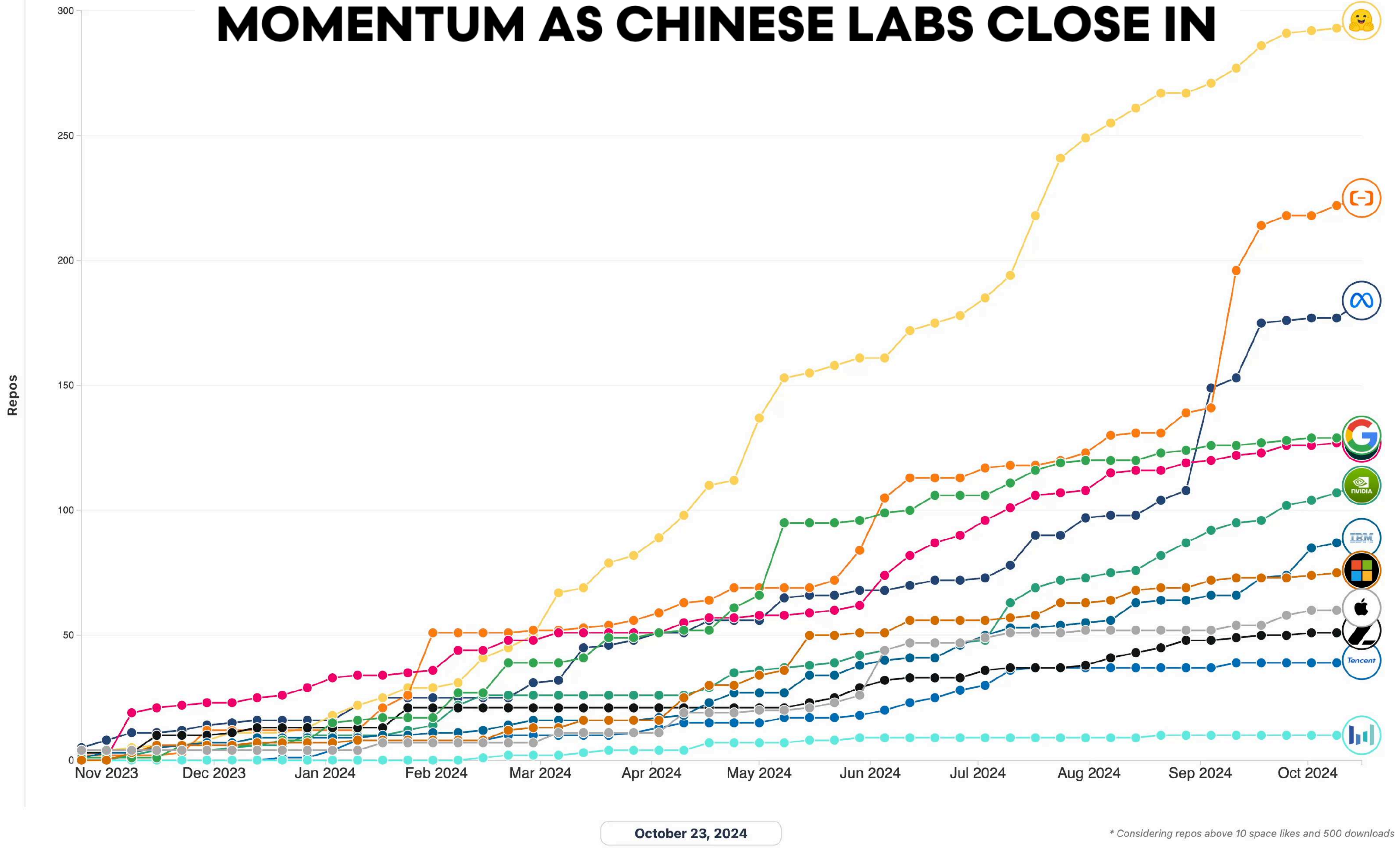
- Nothing is easy in life
- No pain no gain
- But everything is doable with efforts
- The power of collaboration
- Ego is the enemy

Three components to innovate:
Unconventional data 🔥
Unconventional algorithms 🚀
Unconventional collaboration 🌏

Chemistry between the base LLM and RL matters
Conclusions from effortless RL != effortful RL

Conclusions from effortless SFT != effortful SFT https://aiworld.eu/story/nvidia-leads-open-source-ai-momentum-as-chinese-labs-close-in

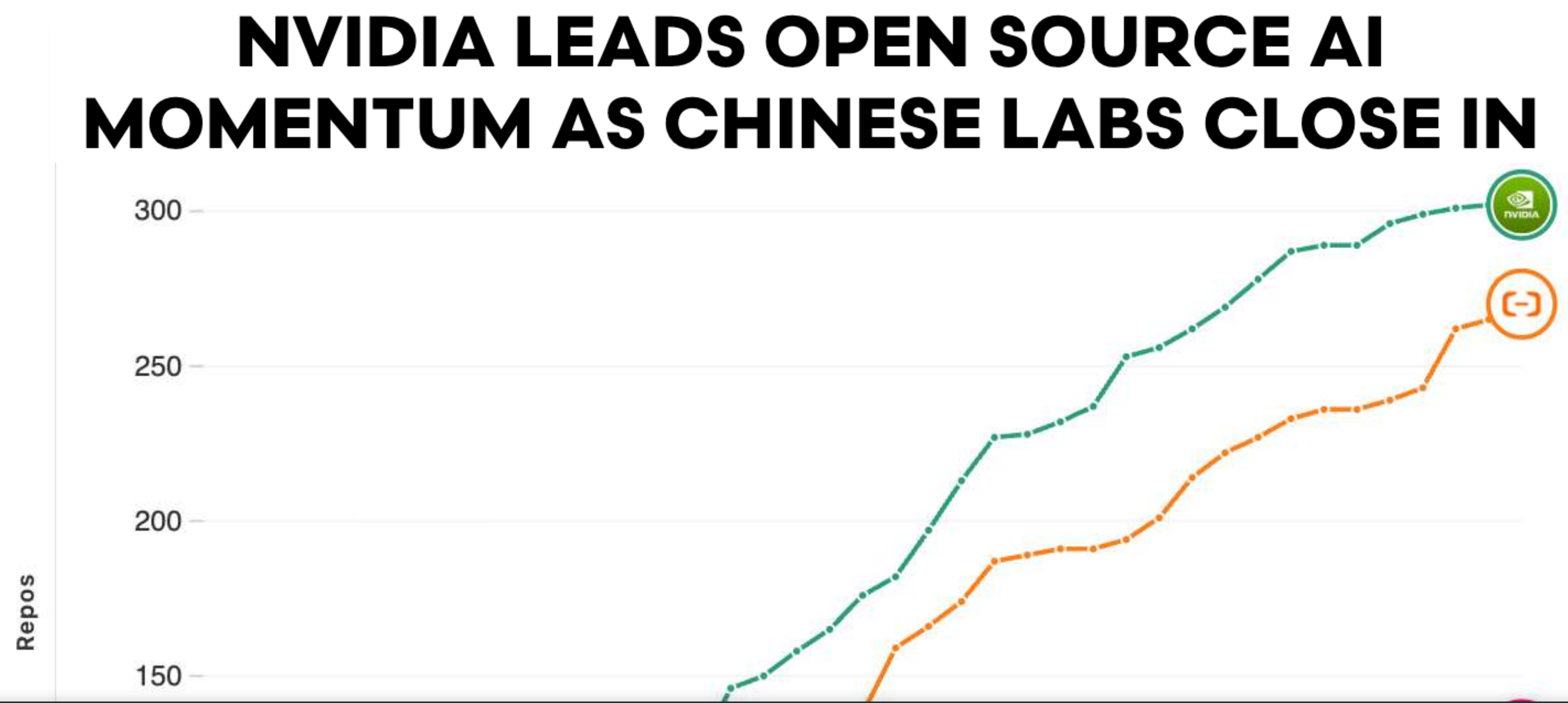
NVIDIA LEADS OPEN SOURCE AI MOMENTUM AS CHINESE LABS CLOSE IN



* Considering repos above 10 space likes and 500 downloads

AI WORLD

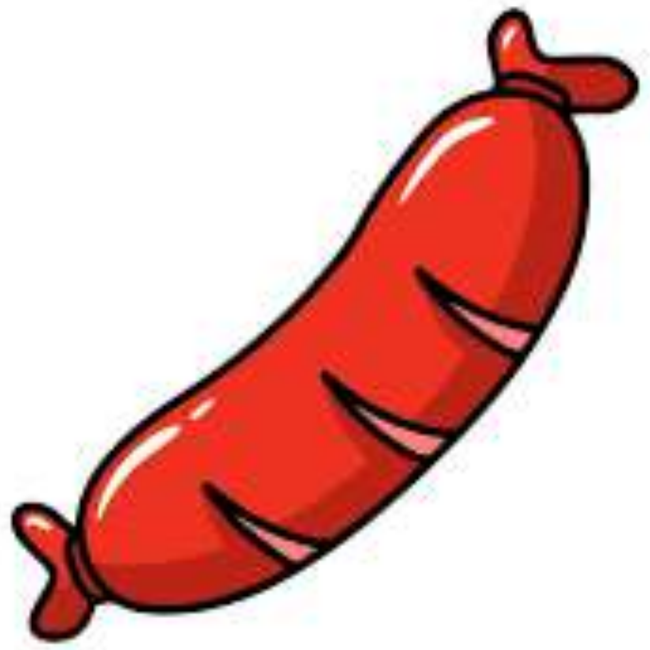
https://aiworld.eu/story/nvidia-leads-open-source-ai-momentum-as-chinese-labs-close-in

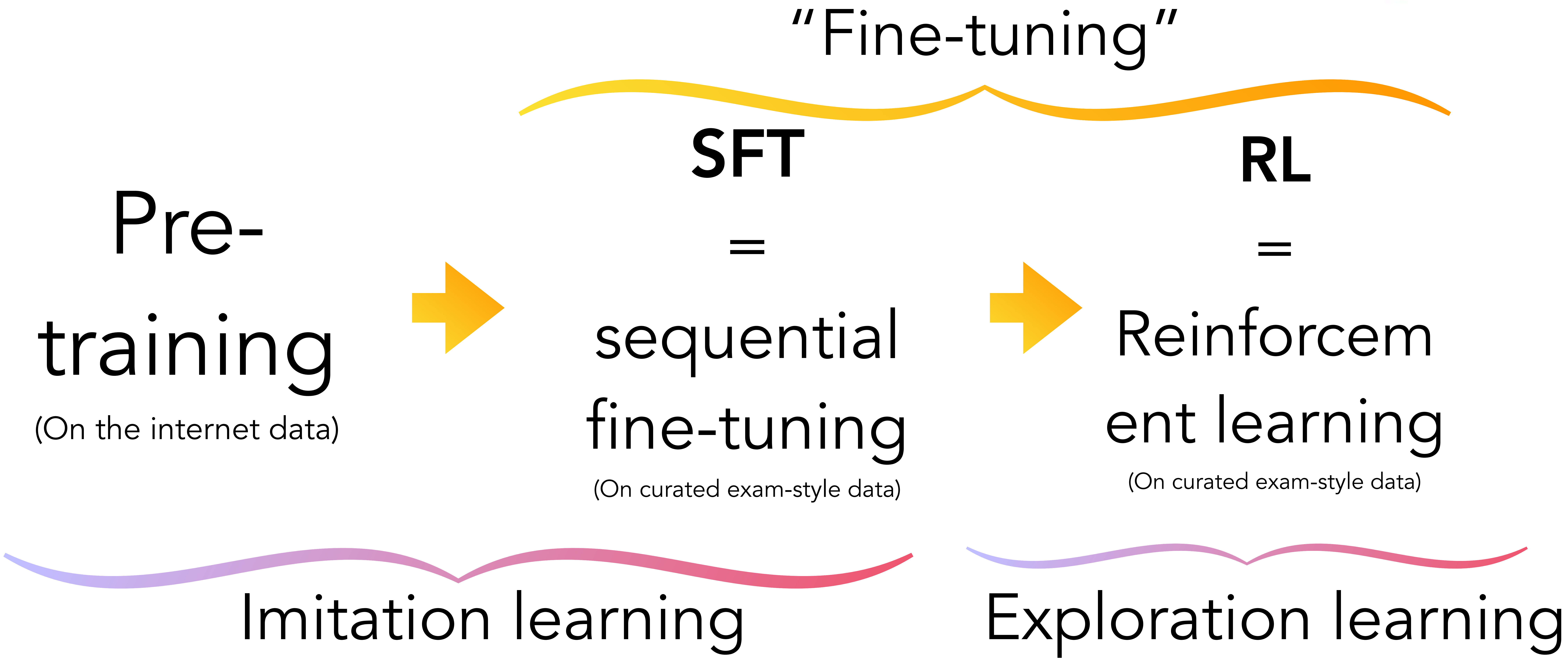


NVIDIA LEADS OPEN SOURCE AI MOMENTUM AS CHINESE LABS CLOSE IN

1. Small models are in high demand, with daily downloads hitting tens of millions (!!! 🔥) for top repositories
2. China's contribution to the open source ecosystem has skyrocketed, with Alibaba now surpassing Hugging Face 😮
3. NVIDIA has emerged as a rising star ⭐ (or a risen star? 😎), for open-source contributions 💚

LLM 101: How sausages are made





"Fine-tuning"

SFT = sequential fine-tuning (On curated exam-style data)

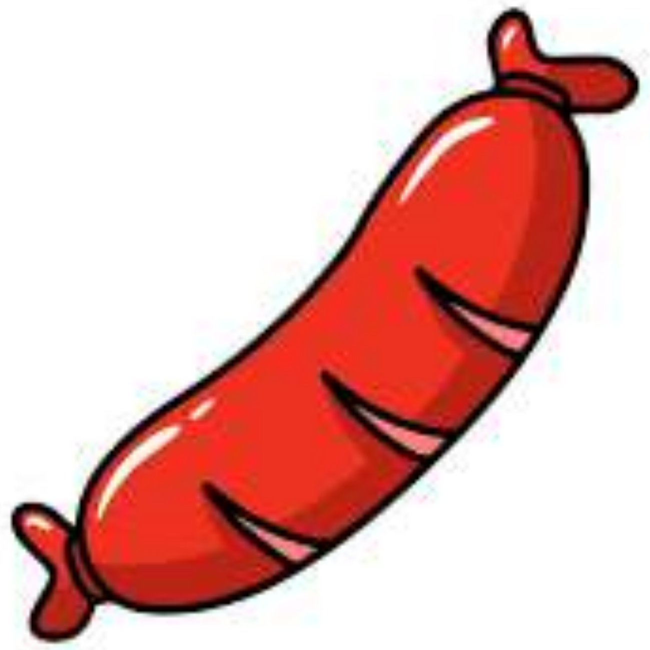
RL = Reinforcement learning (On curated exam-style data)

Pre-training (On the internet data)

Imitation learning

Exploration learning

LLM 101: How sausages are made



"Fine-tuning"

It's all about data

Use entirely of the internet data

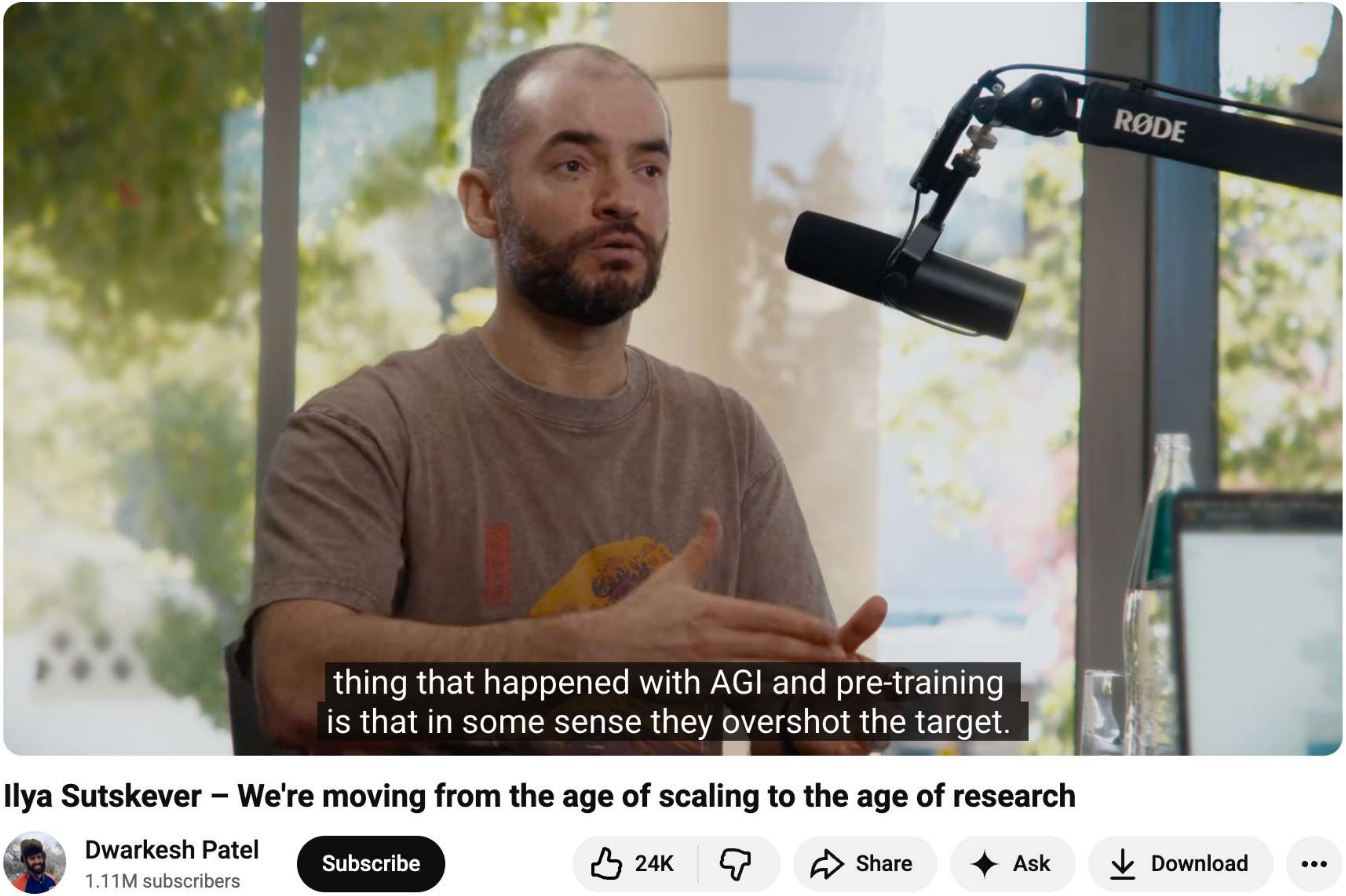
Internet data not enough, so ask humans to write exam data

When even that's not enough, ask AI to synthesize data

=> Make OOD data as in-dist data via brute-force data synth!

Imitation learning

Exploration learning



thing that happened with AGI and pre-training
is that in some sense they overshot the target.

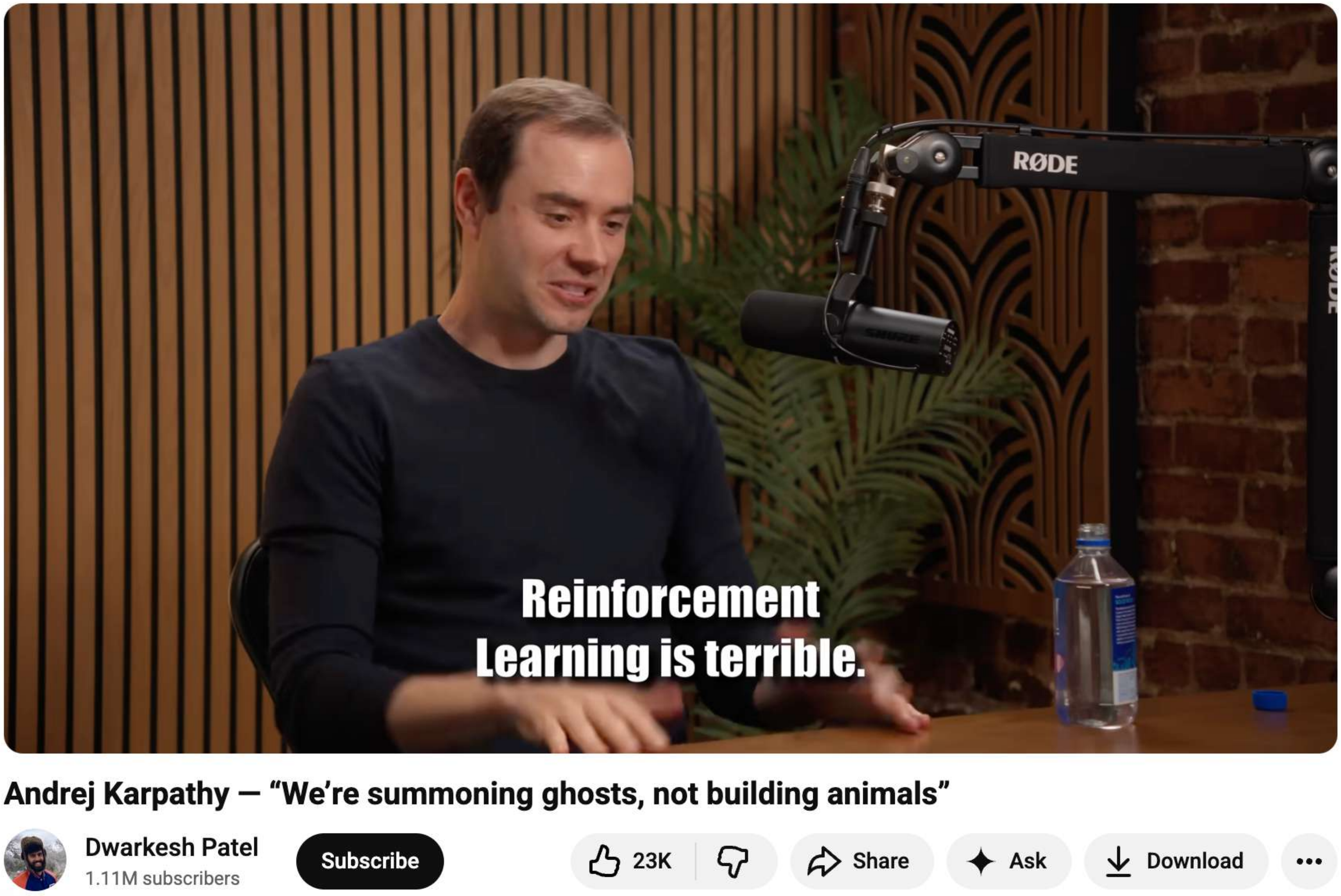
Ilya Sutskever – We're moving from the age of scaling to the age of research

Dwarkesh Patel
1.11M subscribers

Supervised learning is not something that happens in nature.

Richard Sutton – Father of RL thinks LLMs are a dead end

Dwarkesh Patel
1.11M subscribers



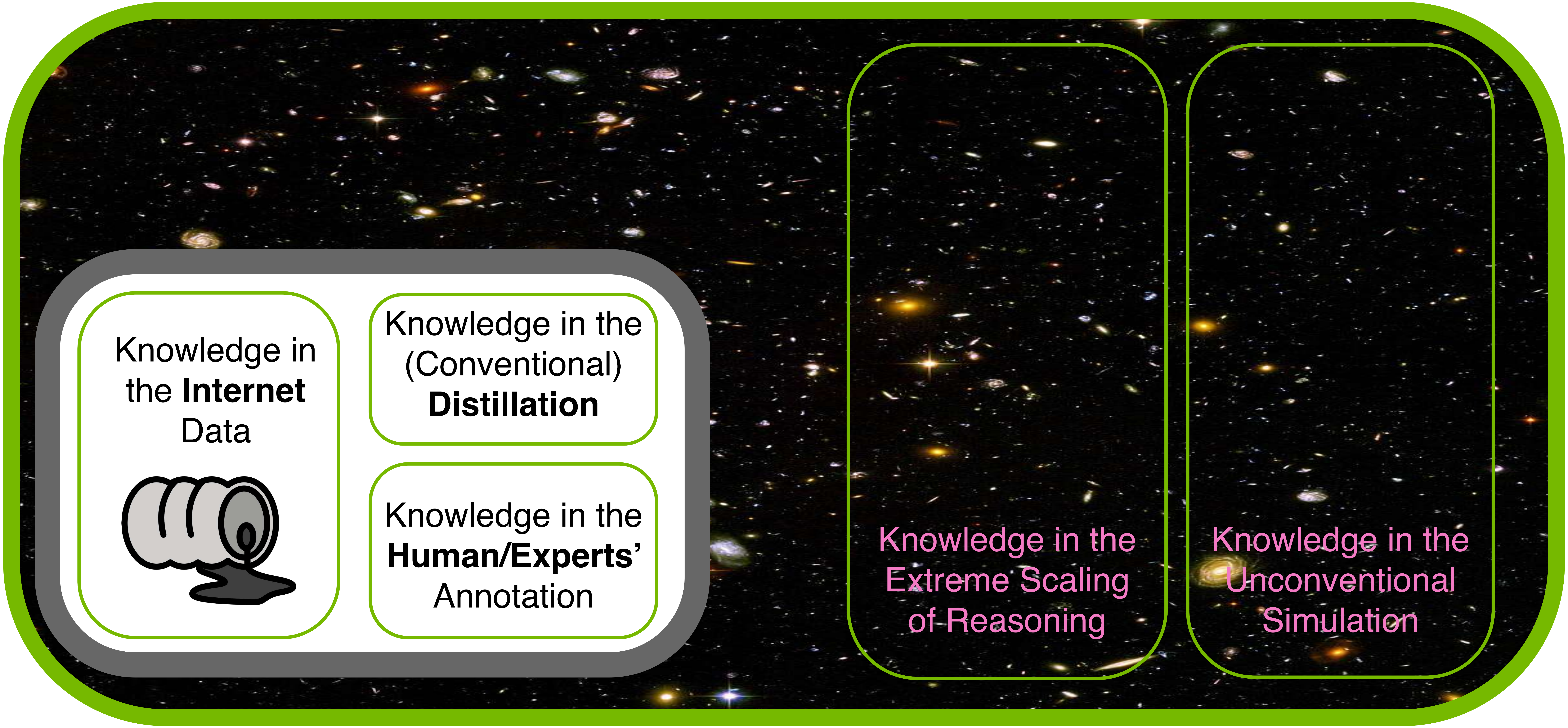
Andrej Karpathy — “We’re summoning ghosts, not building animals”

Dwarkesh Patel

1.11M subscribers

The Universe of Knowledge

= The Universe of Synthetic Data



Knowledge in the **Internet** Data

Knowledge in the (Conventional) **Distillation**

Knowledge in the **Human/Experts'** Annotation

Knowledge in the Extreme Scaling of Reasoning

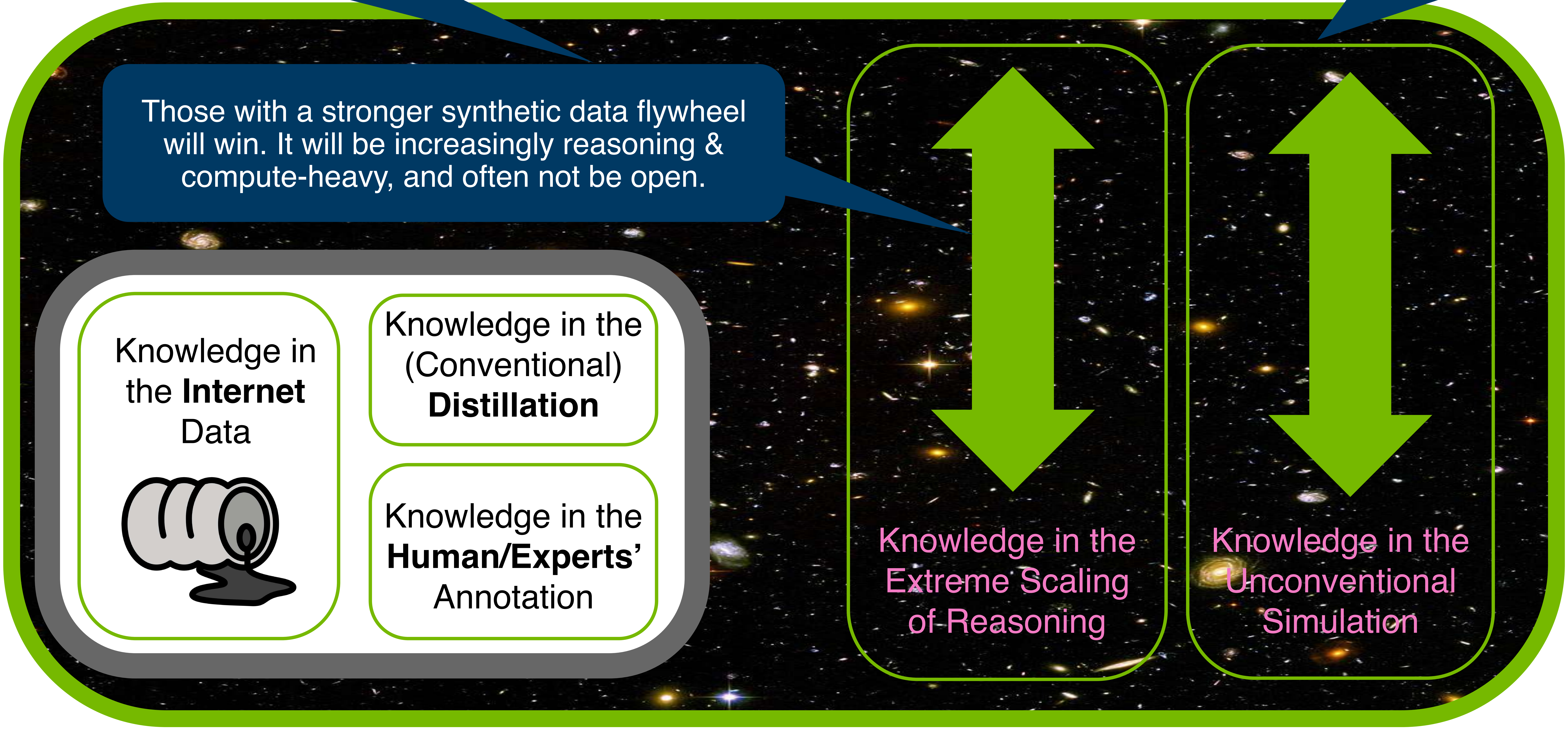
Knowledge in the Unconventional Simulation

Discovery of new knowledge is the act of a scientist, even if the target domain is not for scientific fields per say

iverse of Know

Many real-world use cases require the capability to fill in this data gap

= The Universe of Synthetic Data



Those with a stronger synthetic data flywheel will win. It will be increasingly reasoning & compute-heavy, and often not be open.

Knowledge in the **Internet** Data

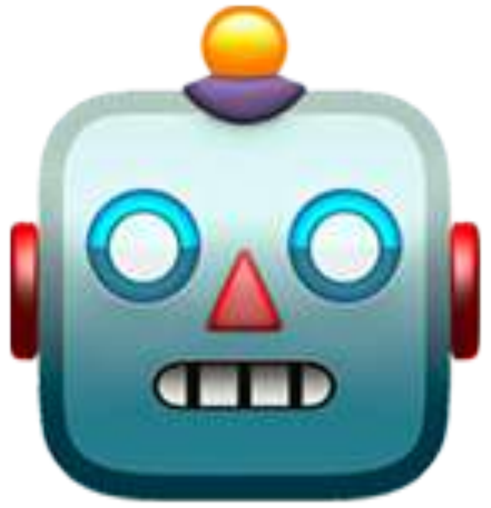
Knowledge in the (Conventional) **Distillation**

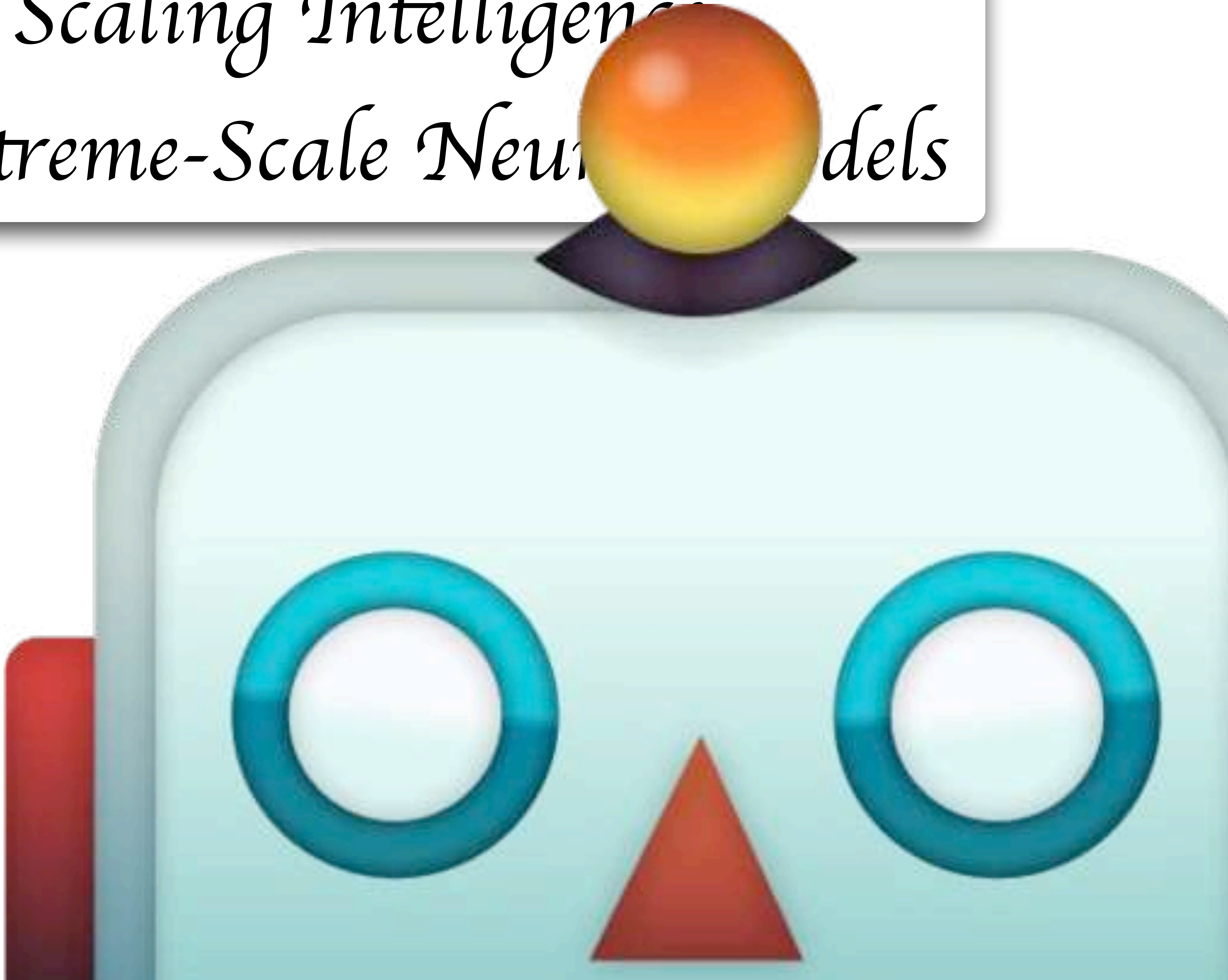
Knowledge in the **Human/Experts'** Annotation

Knowledge in the Extreme Scaling of Reasoning

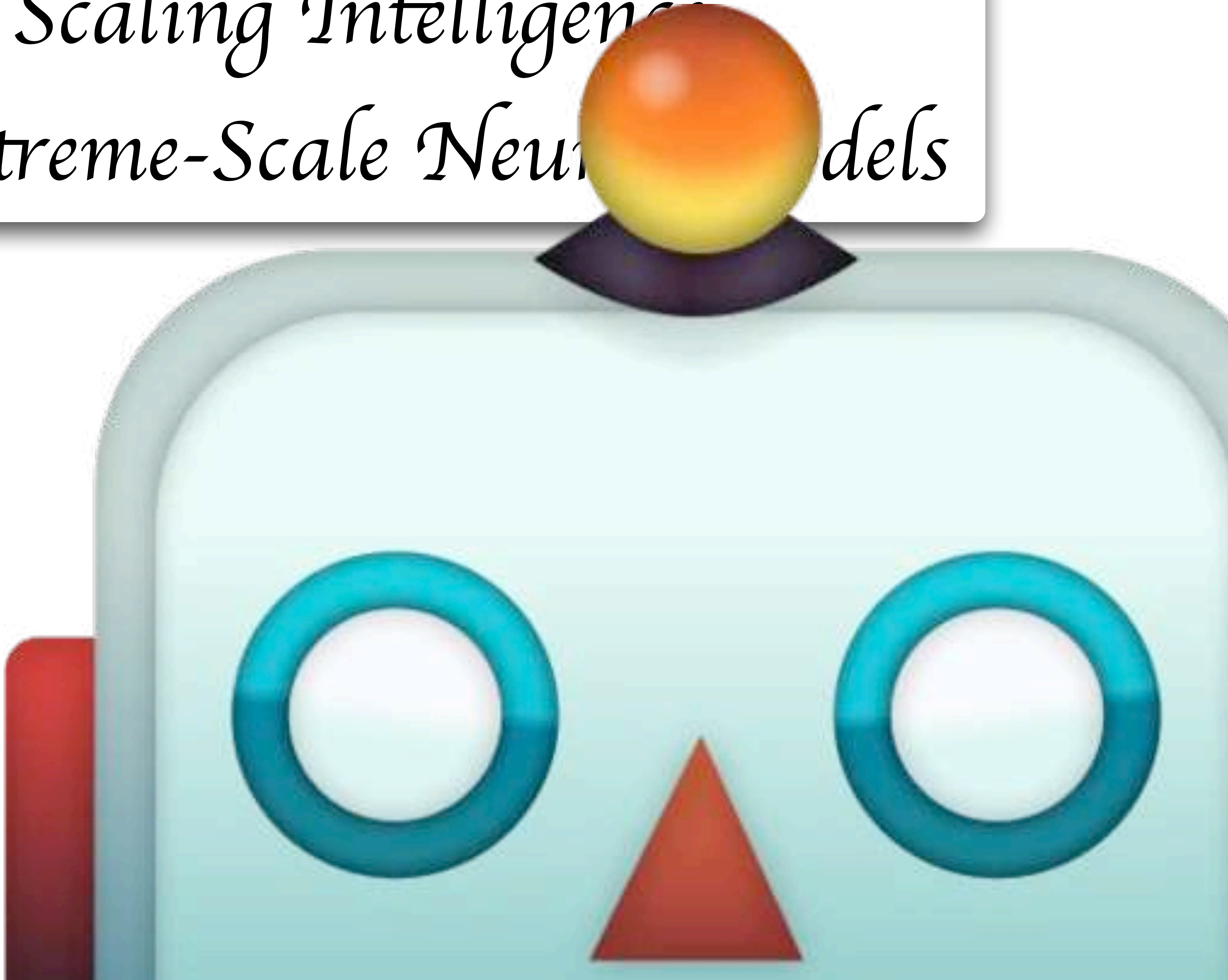
Knowledge in the Unconventional Simulation

David vs. Goliath:
the Art of Scaling Intelligen
in the Era of Extreme-Scale Neu dels





David vs. Goliath:
the Art of Scaling Intelligen
in the Era of Extreme-Scale Neu dels



AI mirrors human intelligence:

- Reasoning is often memorized knowledge
- Exploitation vs exploration trade offs
- Diverse examples/experiences enhance learning

AI diverges from human intelligence:

- More data, compute, and memory
- Less abstraction and conceptualization
- Clear separation between training and testing

How sausages are made:

- 👩🏻‍🚀 Human’s internet data, the artifact of human intelligence
- 👩🏻‍🚀 Human’s audacity to throw in $$$$$$$
- 👩🏻‍🚀 Human’s annotations at scale for alignment
- 👩🏻‍🚀 Human’s intuitions and insights on a lot of engineering details

Open Research Questions

- New theories of intelligences?
- New theories of knowledge and reasoning?
- That humans do not have a context window size of 1M tokens — is this a limitation or blessing?

Thanks!