

DSL Alumni Seminar



통계·데이터사이언스학과 석사 1학기 신재우

Contents

- Before RLVR
- Introduction to RLVR
- 2025 RLVR Trend

Speaker Introduction

신재우

- 통계데이터사이언스학과 석사 1학기 신재우
- 응용통계학과 19.5 학번
- DSL 10기 학술부장
- ESC 32기, 33기 회장
- 손지용 교수님 ITML 연구실 소속
- 연구 분야 : LLM Reasoning + Efficiency



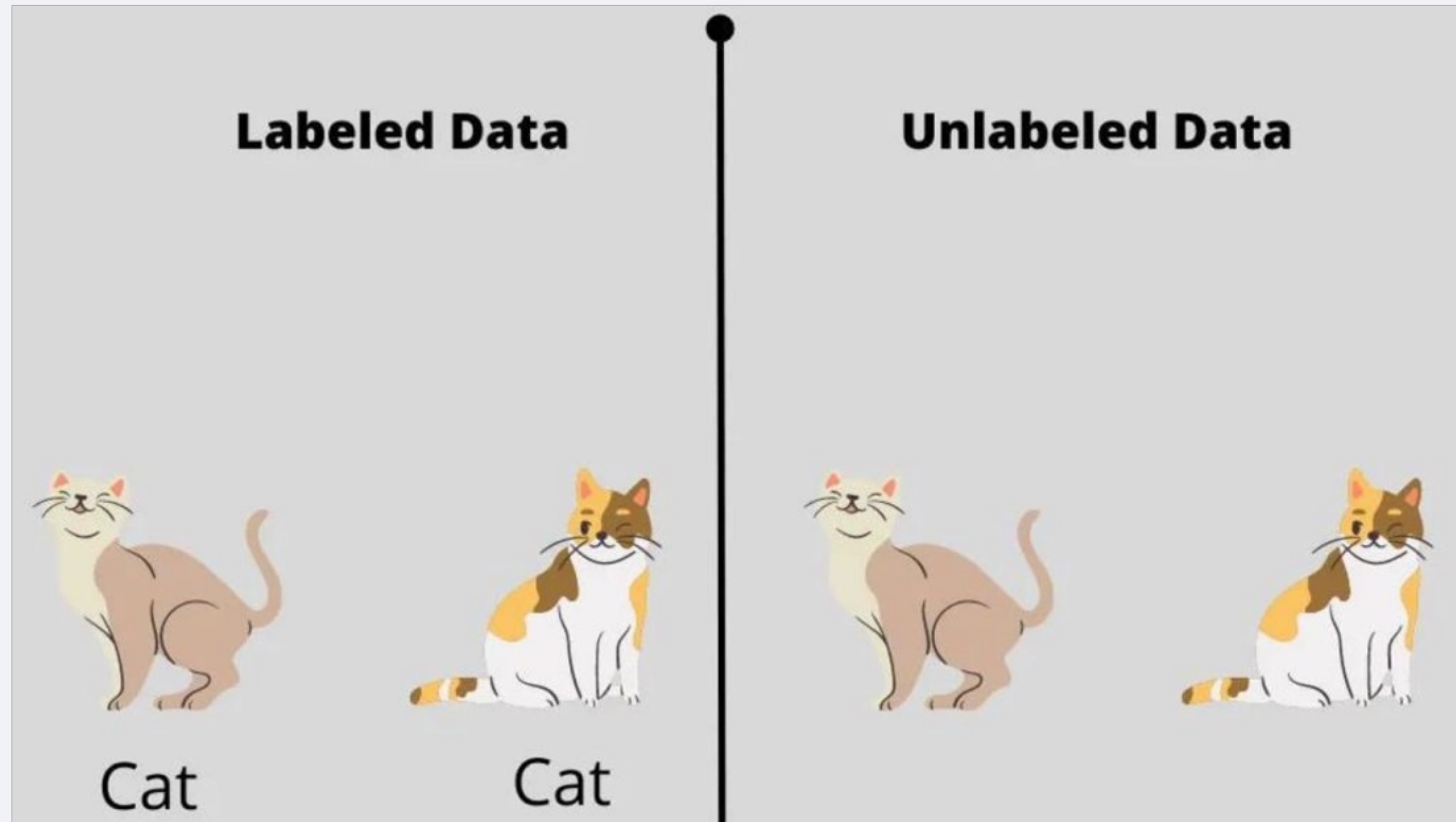
Before RLVR



통계·데이터사이언스학과 석사 1학기 신재우

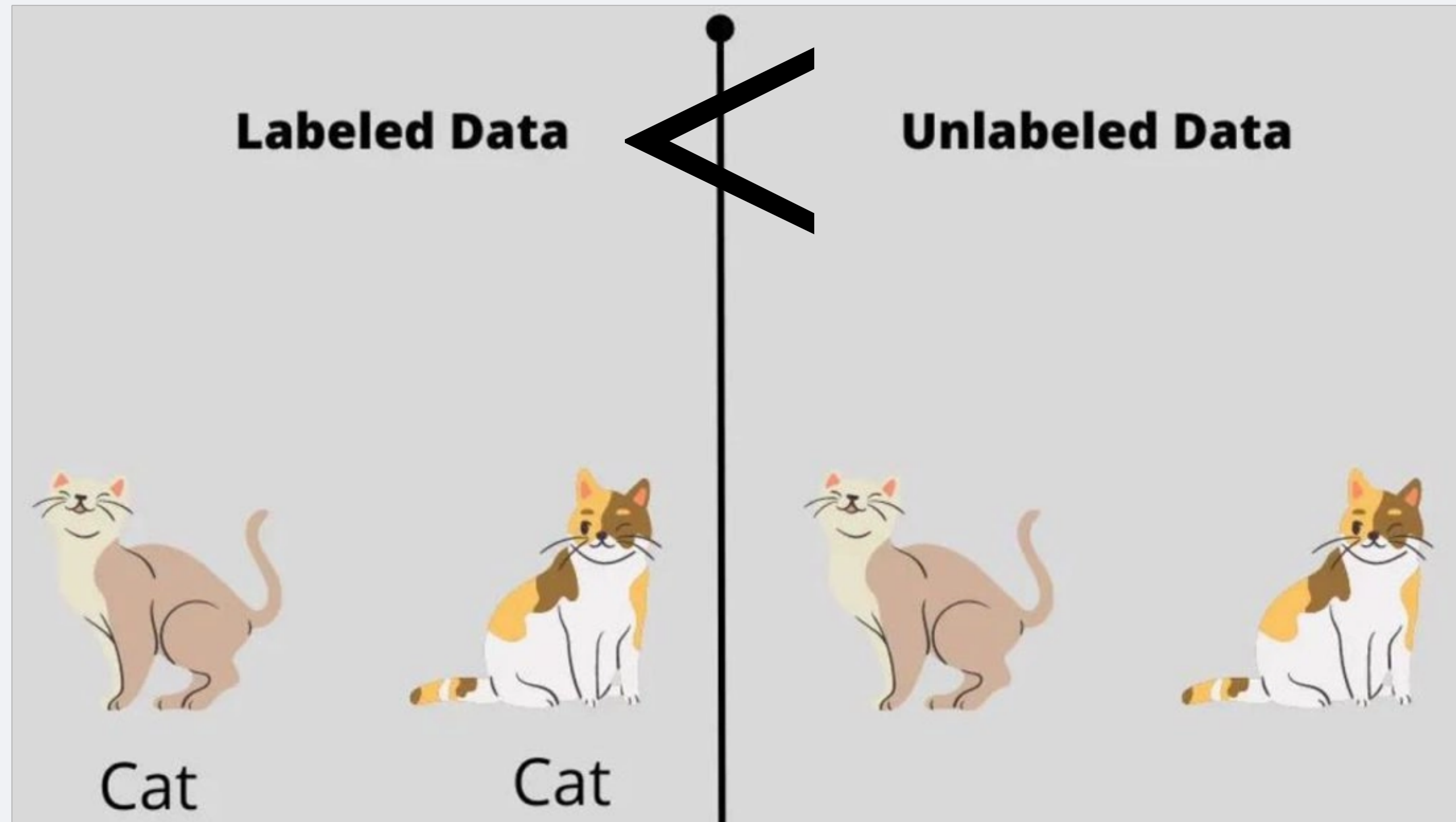
Before RLVR

Which among the two types of data do we have more in this world?



Before RLVR

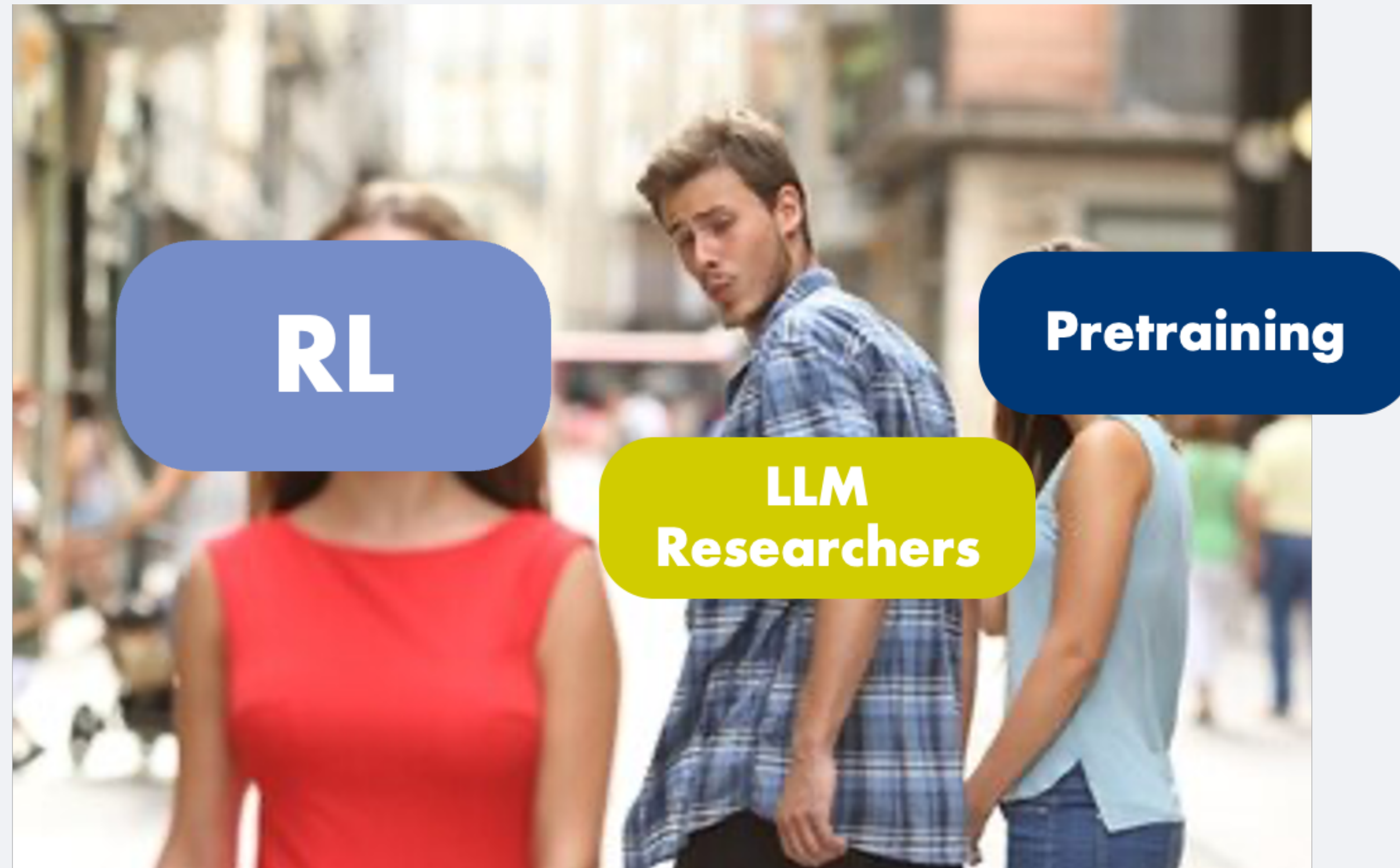
Which among the two types of data do we have more in this world?



Pre-training requires Labeled Data (Texts to follow)

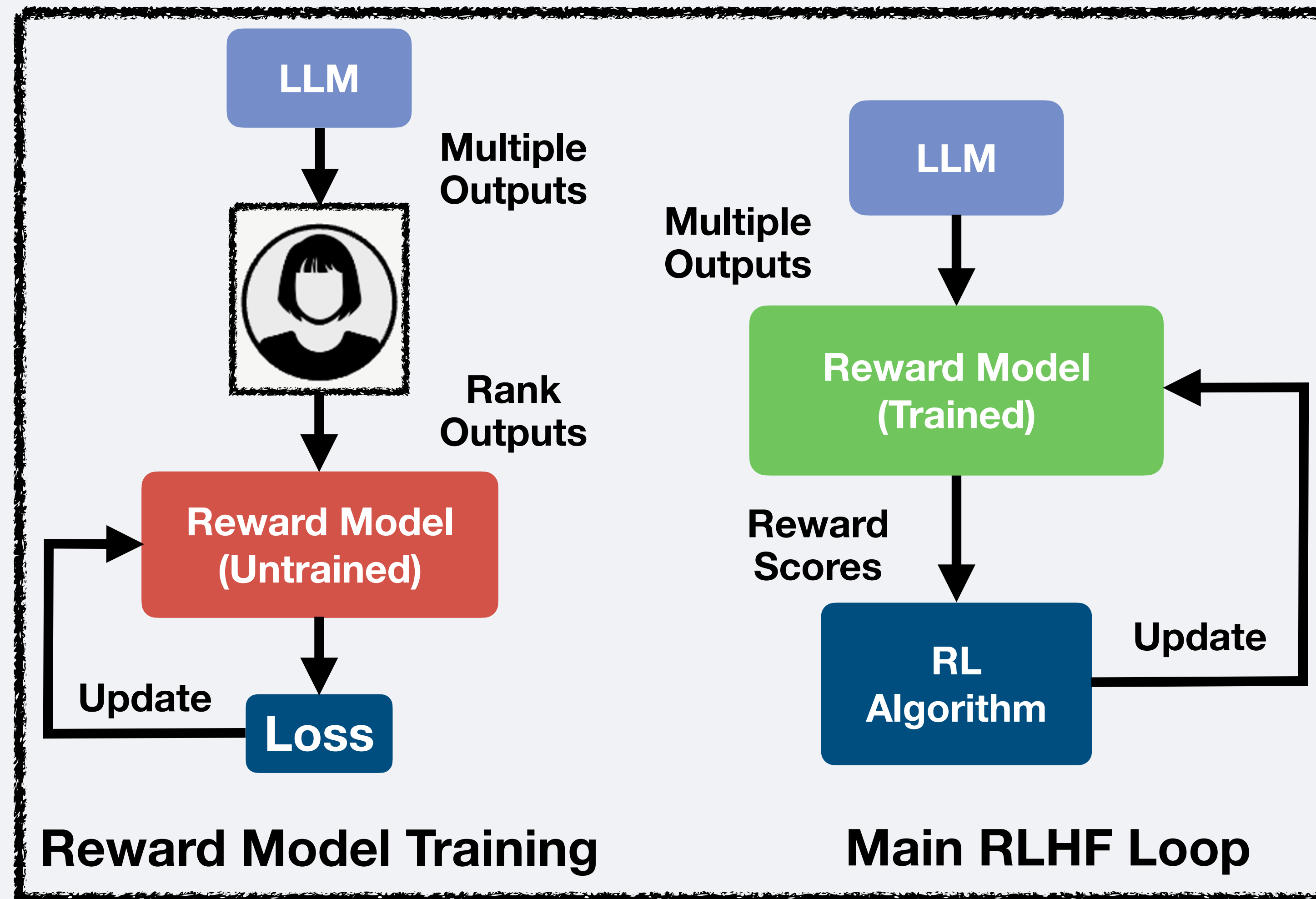
Before RLVR

People started to rely more on RL due to this, as we don't need rationale data.



RLHF

What is **RLHF** (Reinforcement Learning with Human Feedback)?



RLHF

They usually use PPO-style algorithms, as we **don't** want to **deviate** too much from the original model.

Advantage Function

$$\theta_{k+1} = \operatorname{argmax}_{\theta} \mathbb{E}_{(s,a) \sim (\pi_{\theta_k}, T)} \left[\frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)} \overbrace{A^{\pi_{\theta_k}}(s, a)}^{\text{Advantage Function}} \right]$$

such that $\underbrace{\overline{D}_{\text{KL}}(\theta || \theta_k)}_{\text{KL Divergence}} < \delta$

KL Divergence

Clip a value based on an upper and lower bound

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t [\underbrace{\min(r_t(\theta) A_t, \text{CLIP}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_t))}_{\text{Take the minimum of two values}}]$$

Take the minimum of two values

$$\max_{\theta} \mathbb{E} [r_t(\theta) A_t - \beta \text{KL}(\pi_{\theta} || \pi_{\text{ref}})]$$

TRPO Update Rule

Hard Constraint instead of a Penalty.

PPO-CLIP / PPO-KL

Penalty instead of a hard constraint.

RLHF

Summary



**Learns from
the correct
answers
(cross entropy loss)**

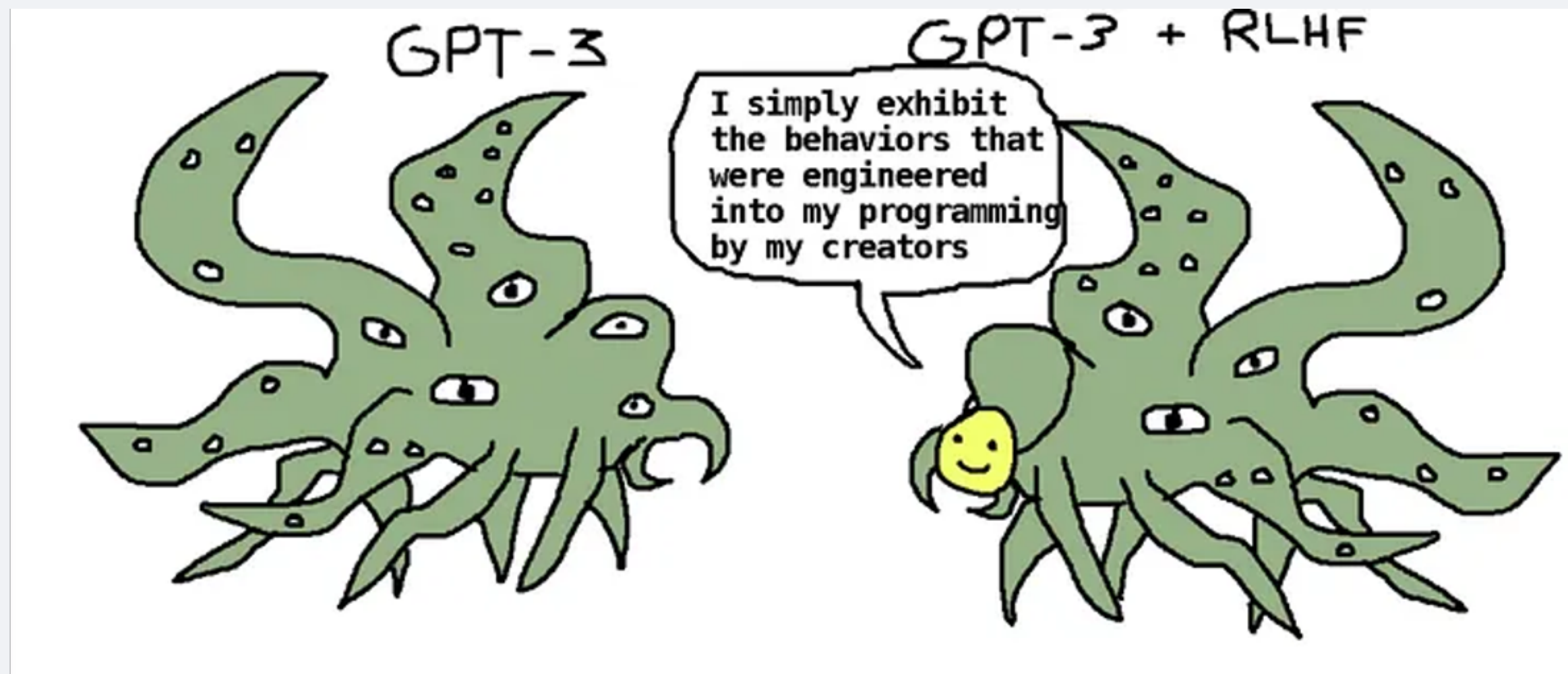
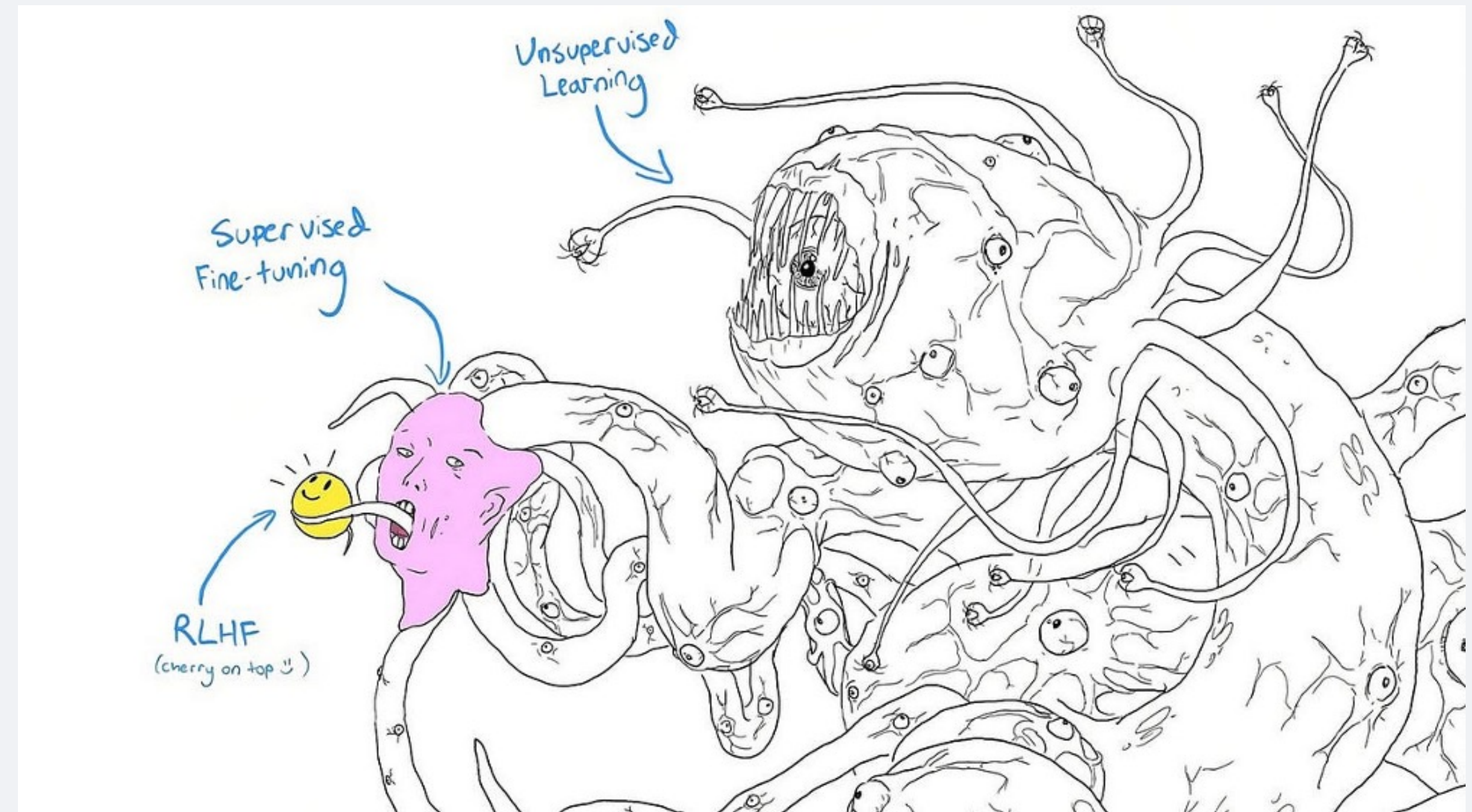


**Learns from
the best
scoring results
(policy loss)**

RLHF and Reasoning

What is the problem of RLHF?

The reward model is trained based off on **preference**, not **Correct / Wrong** standards.



RLHF and Reasoning

Question we want to answer :

Does this Preference-based learning actually have no effect on Reasoning?

How do we test this?

Reasoning Tasks are mostly **Symbolic / Visual**, so we test them by changing these tasks to **Text**.

Large Language Models Are Not Strong Abstract Reasoners

Gaël Gendron

University of Auckland
ggen187@aucklanduni.ac.nz

Qiming Bao

University of Auckland
qbao775@aucklanduni.ac.nz

Michael Witbrock

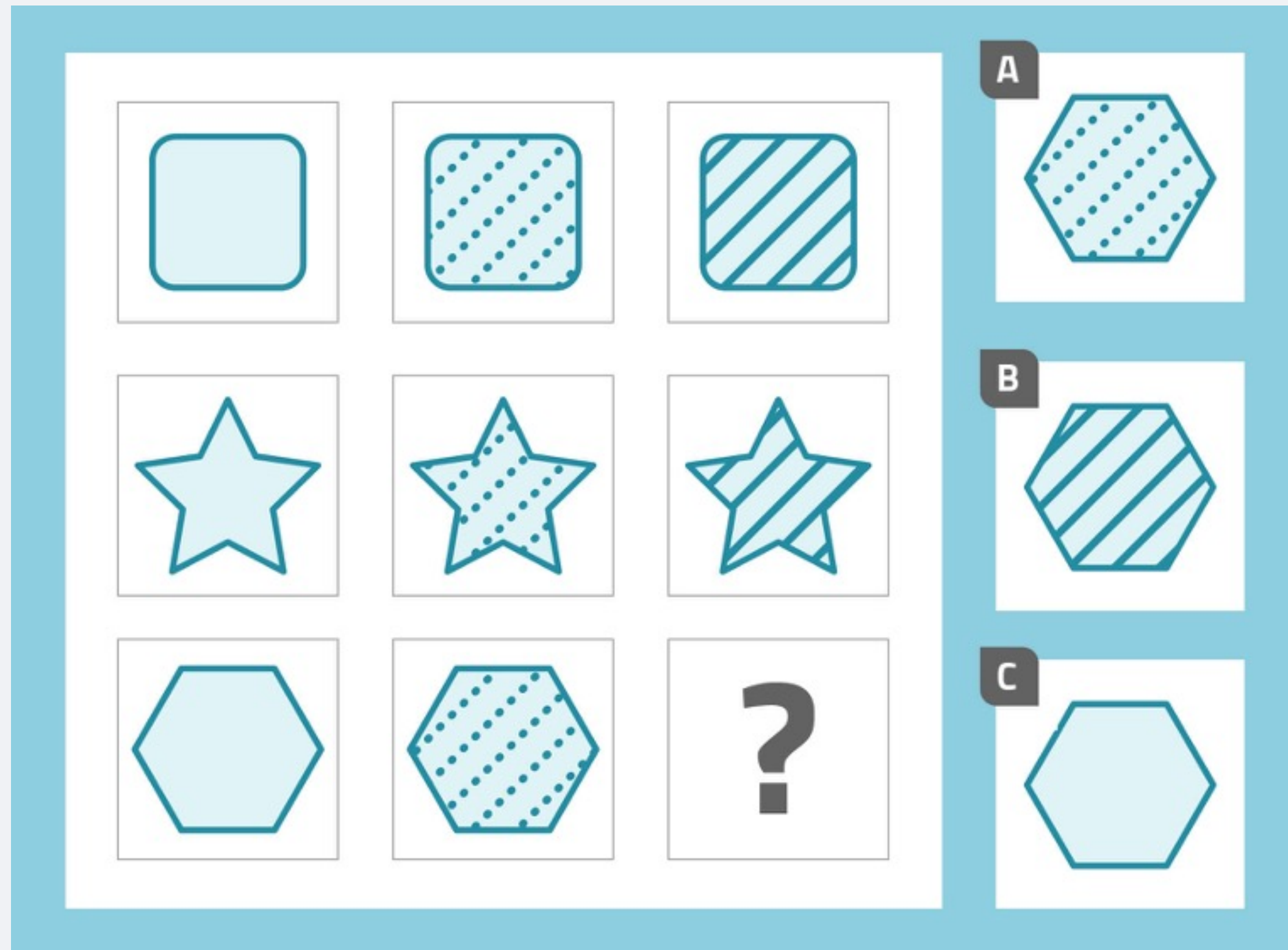
University of Auckland
m.witbrock@auckland.ac.nz

Gillian Dobbie

University of Auckland
g.dobbie@auckland.ac.nz

RLHF and Reasoning

Example — Raven's Progressive Matrices



Example Cases

1. On an image, a large lime square rotated at 180 degrees.
2. On an image, a medium lime square rotated at 180 degrees.
3. On an image, a huge lime square rotated at 180 degrees.
4. On an image, a huge yellow circle rotated at 0 degrees.
5. On an image, a large yellow circle rotated at 0 degrees.
6. On an image, a medium yellow circle rotated at 0 degrees.
7. On an image, a medium white hexagon rotated at -90 degrees.
8. On an image, a huge white hexagon rotated at -90 degrees.

- A. On an image, a tiny white hexagon rotated at -90 degrees.
- B. On an image, a giant white hexagon rotated at -90 degrees.
- C. On an image, a large red hexagon rotated at -90 degrees.
- D. On an image, a large orange hexagon rotated at -90 degrees.
- E. On an image, a large white hexagon rotated at -90 degrees.
- F. On an image, a large green hexagon rotated at -90 degrees.
- G. On an image, a large blue hexagon rotated at -90 degrees.
- H. On an image, a large yellow hexagon rotated at -90 degrees.

RLHF and Reasoning

Results on Multiple Choice Datasets

+ RLHF, >100B

SFT only, <10B

Table 4: Accuracy of Large Language Models for Multiple-Choice QA on the ACRE^T, Evals-P and RAVEN^T datasets. The last line indicates random performance. Completion models can perform worse than random if they do not reply with a valid answer. The best result for each dataset is indicated in **bold**, and the second best is indicated in *italics*.

	ACRE ^T		Evals-P	RAVEN ^T -mcqa	
	Text	Symb		Text	Symb
GPT-2	0.371	0.00	0.496	0.00	0.126
Text-Davinci-3	0.098	0.427	<i>0.560</i>	<i>0.461</i>	<i>0.452</i>
GPT-3.5-Turbo	0.184	0.445	0.481	0.276	0.315
GPT-4	<i>0.272</i>	<i>0.512</i>	0.625	0.697	0.535
LLaMA-7B	0.000	0.257	0.544	0.004	0.000
LLaMA2-7B	0.014	0.003	0.500	0.026	0.149
Alpaca	0.036	0.238	0.544	0.015	0.058
Alpaca-LoRA	0.015	0.123	0.552	0.082	0.124
Zephyr-7B-β	0.106	0.516	0.504	0.000	0.022
random	0.33	0.33	0.5	0.125	0.125

RLHF and Reasoning

RLHF is not really the key for raising Reasoning Capabilities of LLM's!

We need something else...

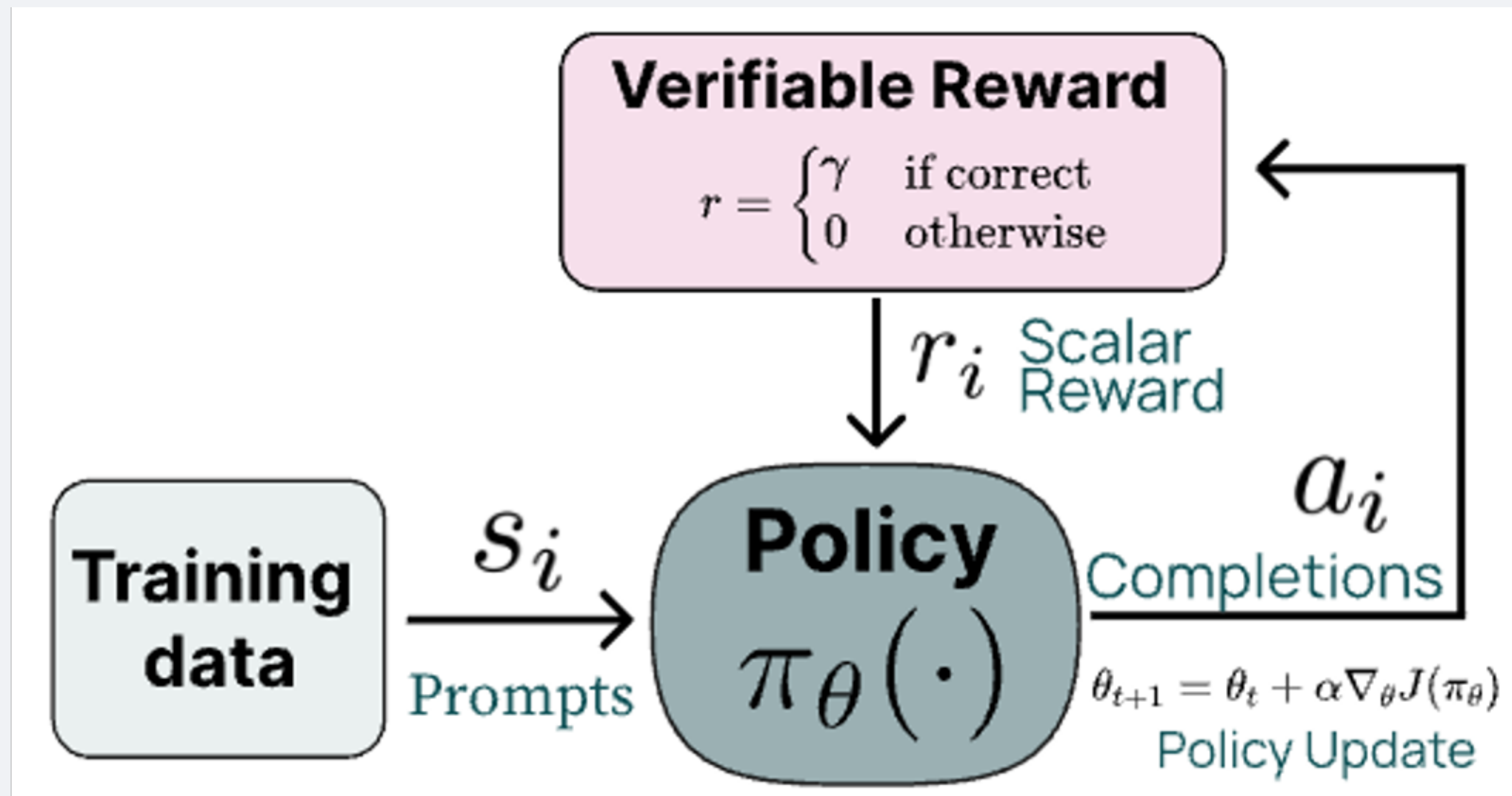
Introduction to RLVR



통계·데이터사이언스학과 석사 1학기 신재우

Introduction to RLVR

What was **RLVR** again...?



Introduction to RLVR

What was RLVR again...?

Problem

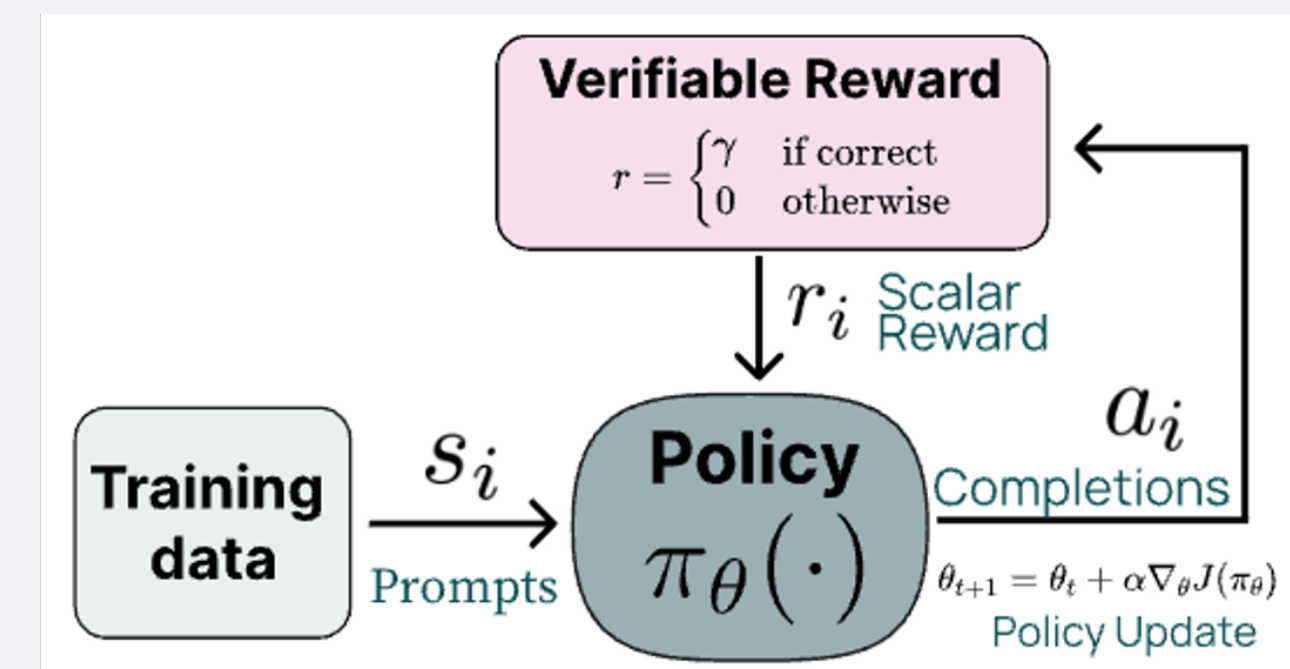
The battery charge in Mary's cordless vacuum cleaner lasts ten minutes. It takes her four minutes to vacuum each room in her house. Mary has three bedrooms, a kitchen, and a living room. How many times does Mary need to charge her vacuum cleaner to vacuum her whole house?

Solution

Mary has $3 + 1 + 1 = 5$ rooms in her house.
At 4 minutes a room, it will take her $4 * 5 = 20$ minutes to vacuum her whole house.
At 10 minutes a charge, she will need to charge her vacuum cleaner $20 / 10 = 2$ times to vacuum her whole house.

Final Answer

2



Input : x_i

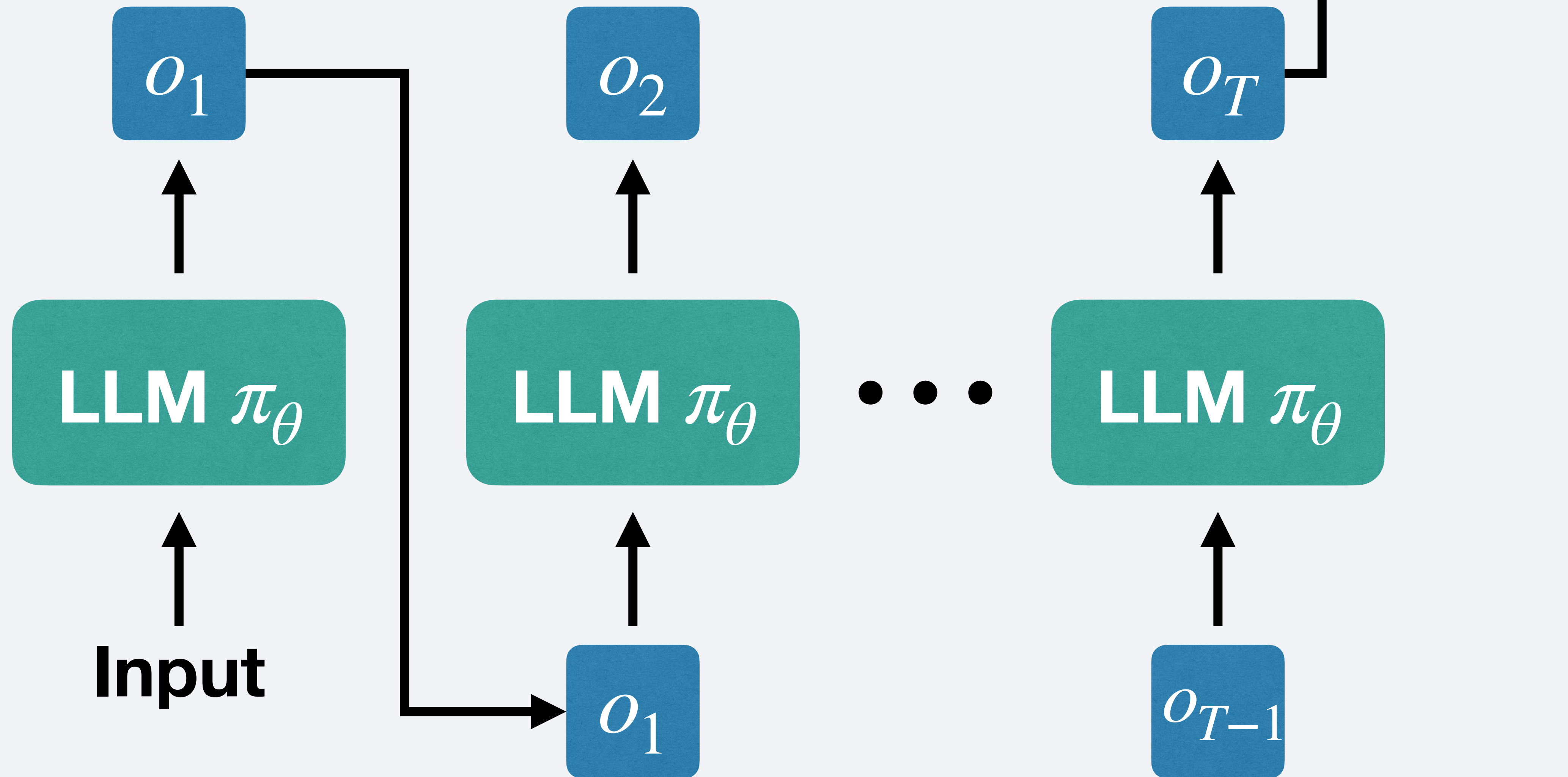
Rationale : $o_{<T}$

Final Answer : o_T

$$r_i = \mathbf{1} \left(o_T = y_i \right)$$

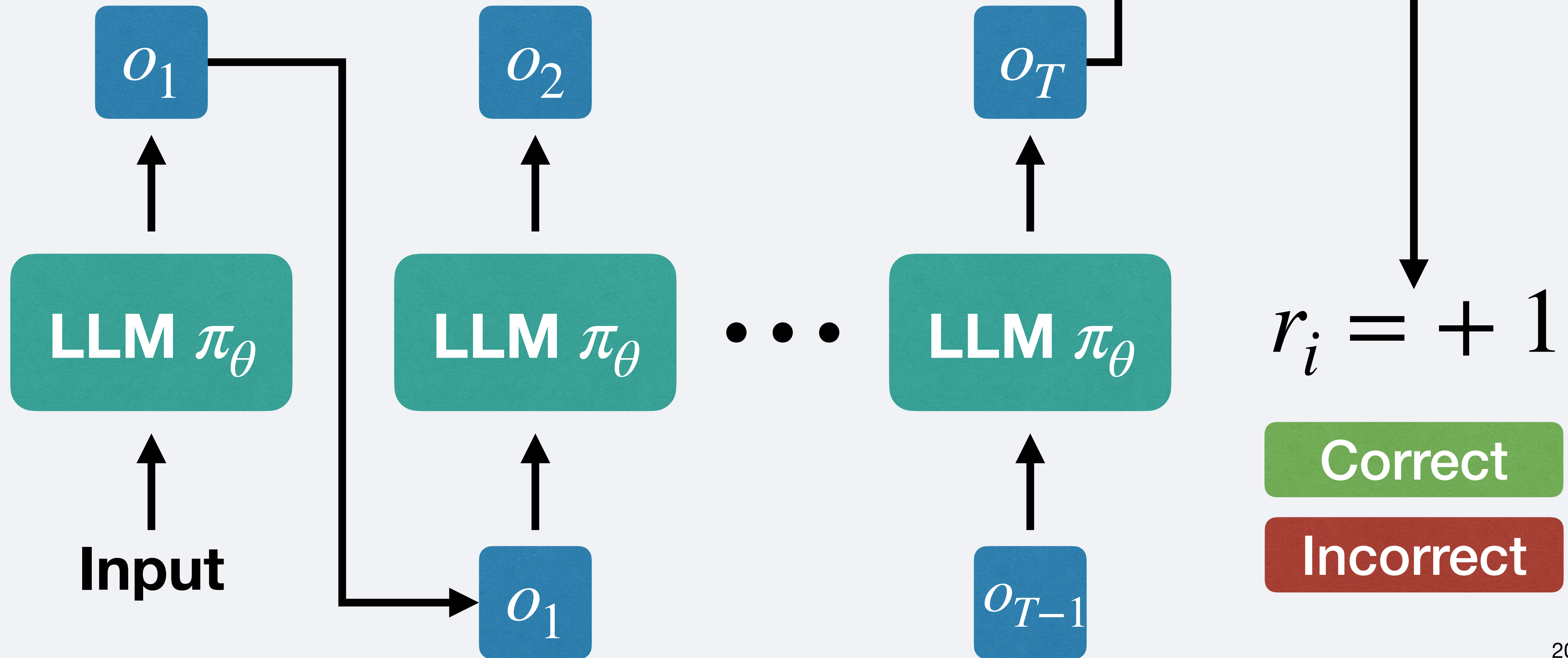
Introduction to RLVR

How does RLVR work again?



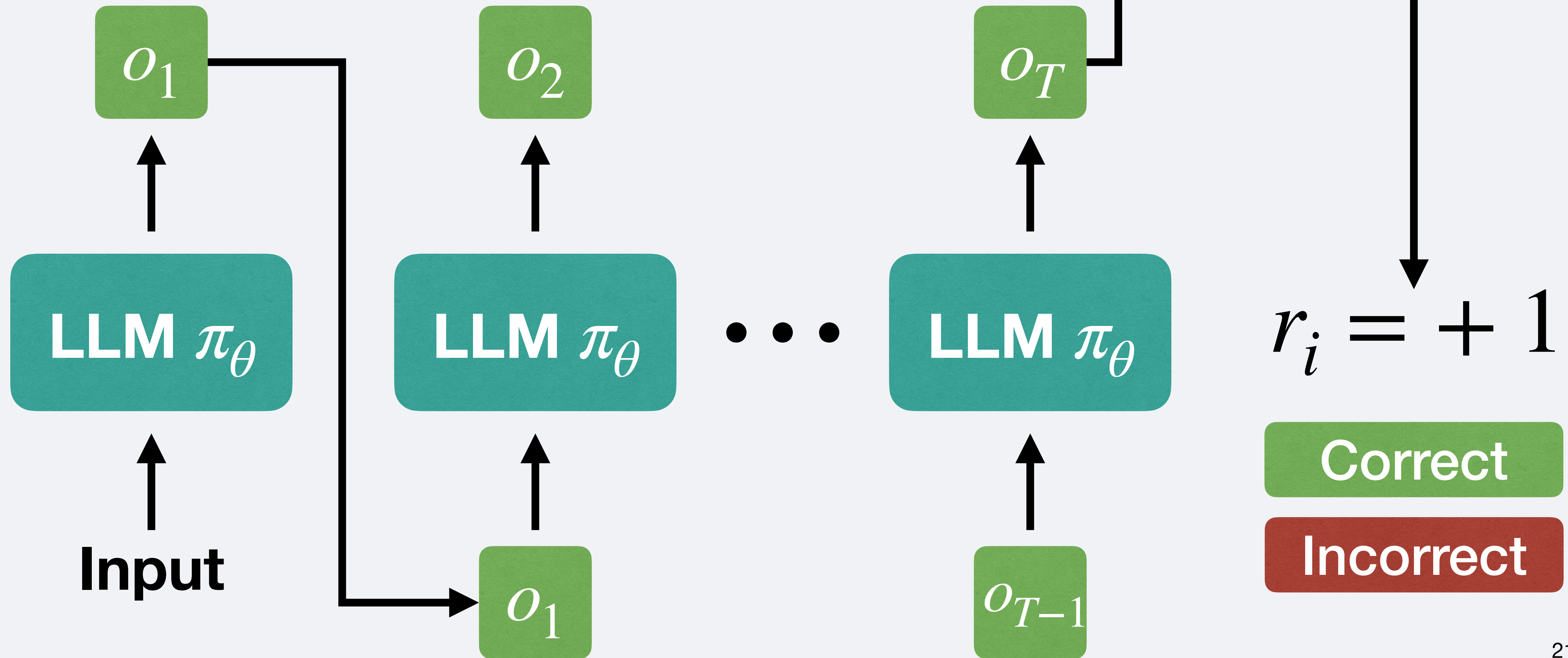
Introduction to RLVR

How does RLVR work again?



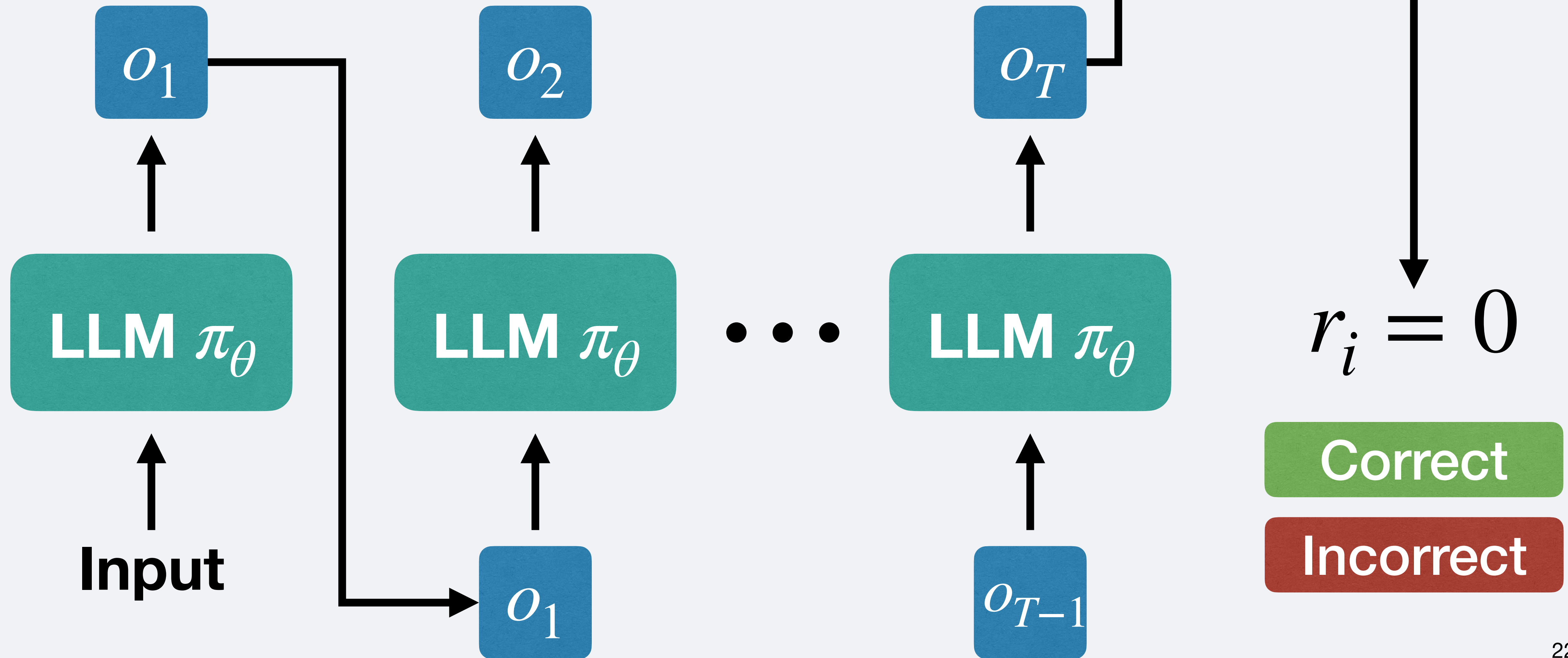
Introduction to RLVR

How does RLVR work again?



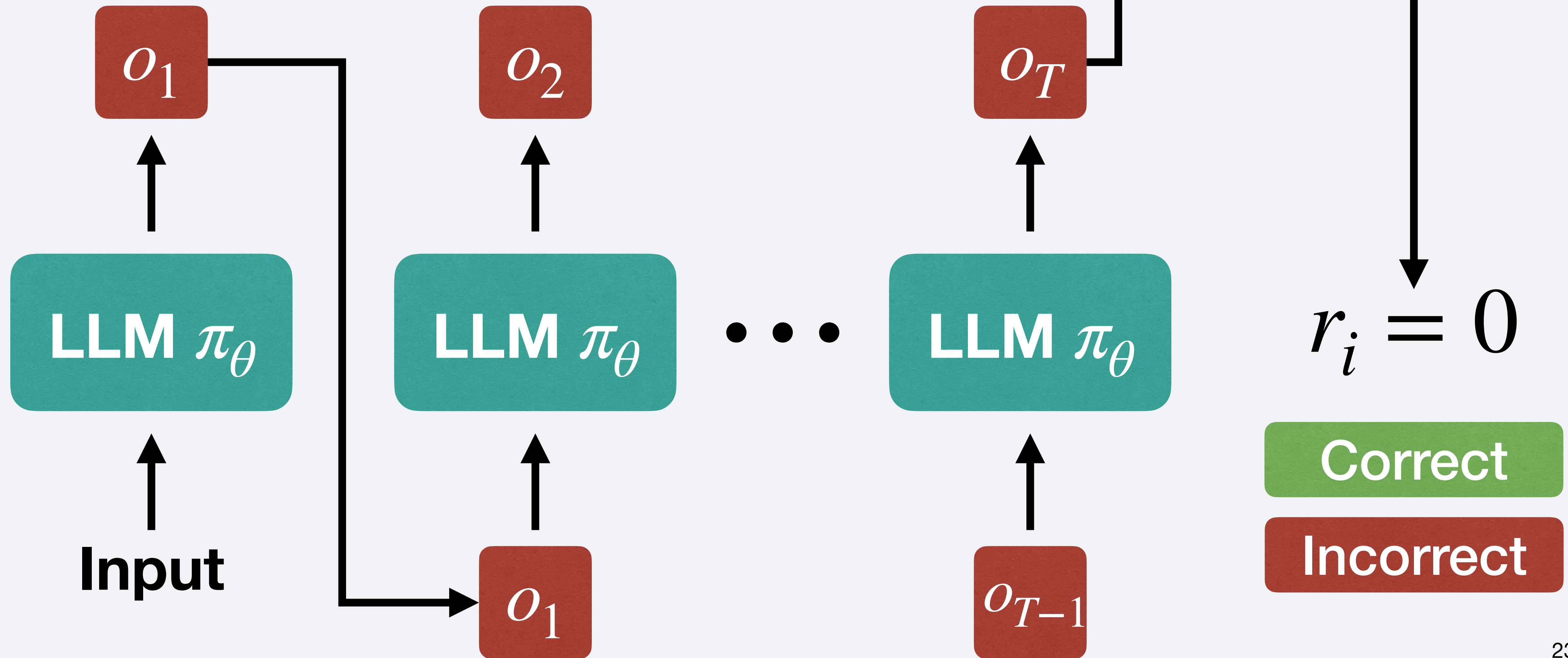
Introduction to RLVR

How does RLVR work again?



Introduction to RLVR

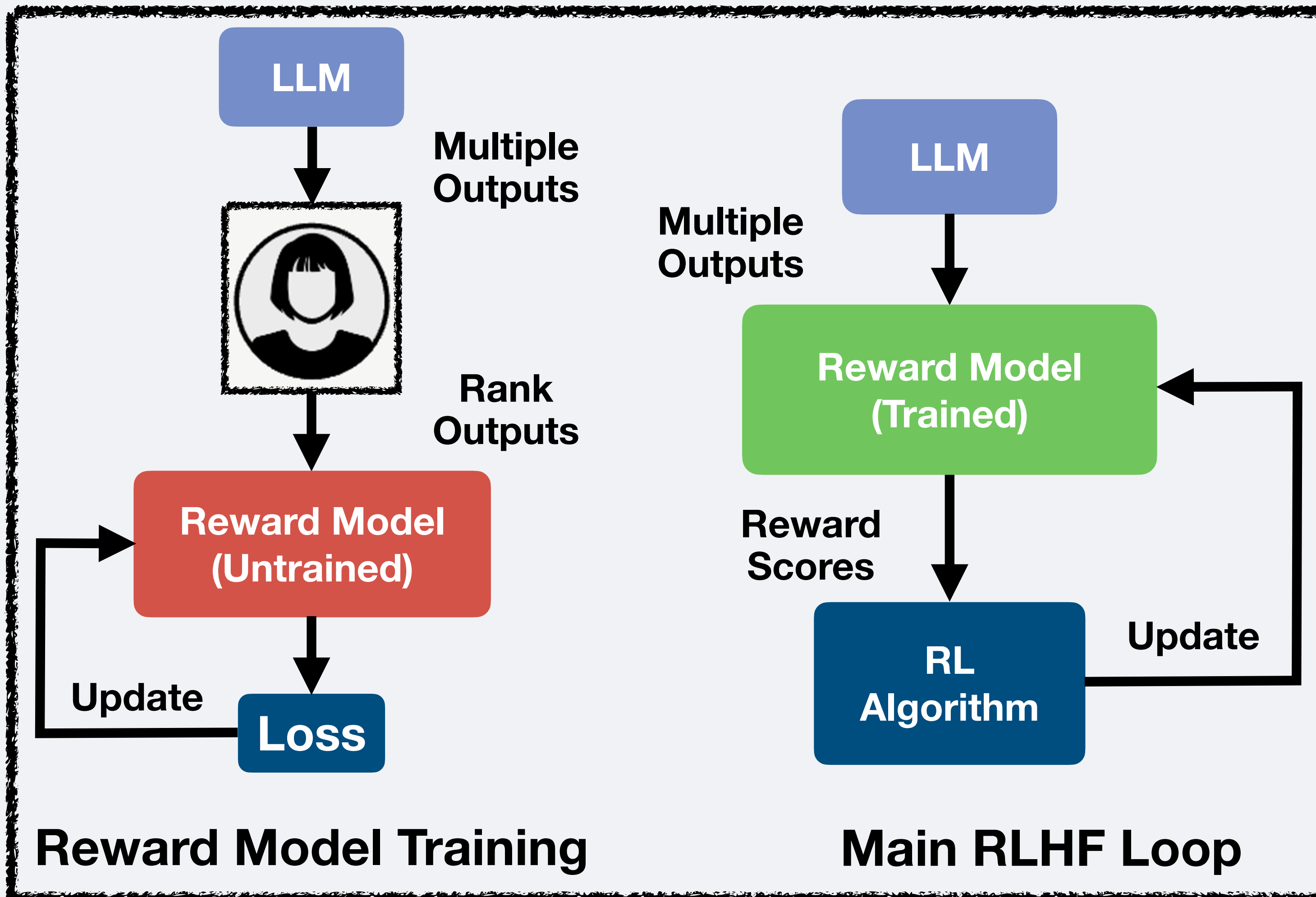
How does RLVR work again?



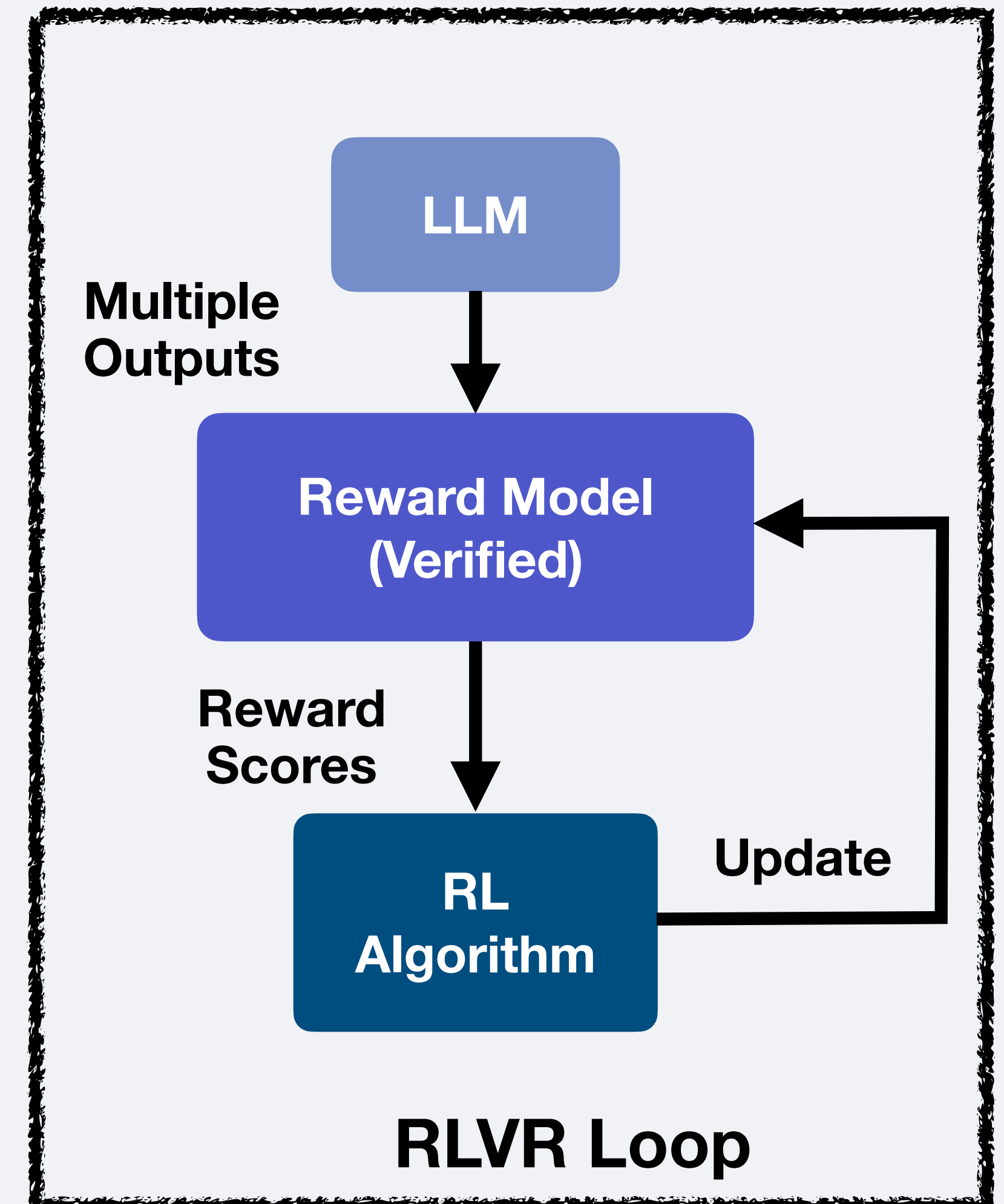
Introduction to RLVR

What is the difference between RLVR and RLHF?

RLHF



RLVR



Tulu 3

The first paper to introduce the concept of RLVR is actually Tulu-3.



Tulu 3: Pushing Frontiers in Open Language Model Post-Training

Nathan Lambert^{1,*} Jacob Morrison¹ Valentina Pyatkin^{1,2} Shengyi Huang¹ Hamish Ivison^{1,2}
Faeze Brahman¹ Lester James V. Miranda¹

Alisa Liu² Nouha Dziri¹ Xinxu Lyu¹ Yuling Gu¹ Saumya Malik¹ Victoria Graf² Jena D. Hwang¹
Jiangjiang Yang¹ Ronan Le Bras¹ Oyvind Tafjord¹ Chris Wilhelm¹

Luca Soldaini¹ Noah A. Smith^{1,2} Yizhong Wang^{1,2} Pradeep Dasigi¹ Hannaneh Hajishirzi^{1,2}
¹Allen Institute for AI, ²University of Washington

**PPO + Soft KL-
Penalty**

$$\max_{\theta} \mathbb{E} [r_t(\theta) A_t - \beta \text{KL}(\pi_{\theta} || \pi_{\text{ref}})]$$

Tulu 3

What is A_t here? → **Advantage Term**

**PPO + Soft KL-
Penalty**

$$\max_{\theta} \mathbb{E} [r_t(\theta) A_t - \beta \text{KL}(\pi_{\theta} || \pi_{\text{ref}})]$$

Advantage Function

$$A(s, a) = \underbrace{Q(s, a)}_{\text{q value for action a in state s}} - \underbrace{V(s)}_{\text{average value of that state}}$$

q value for action a
in state s

average
value
of that
state

The problem for Tulu 3 is that we need to **Train the Value Model**.

- **RLVR Training:** For inference, we deployed the model using vLLM with 16-way tensor parallelism, while utilizing the remaining 240 GPUs for training. After each iteration of RLVR update, the weights are synchronized to the vLLM engine using NCCL broadcast. Inference typically takes ~550 seconds, weight transfer takes ~25 seconds, and training takes ~1,500 seconds. To reduce computational cost during the RLVR stage, we utilized an 8B value model. Future works can benefit from exploring larger value models or alternate value model-free RL algorithms such as GRPO (Shao et al., 2024).

DeepSeek

Suddenly, then DeepSeek was a thing, where they utilized Group-Relative Policy Optimization (**GRPO**) with RLVR.

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

`research@deepseek.com`

DeepSeek

What was GRPO again...?

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

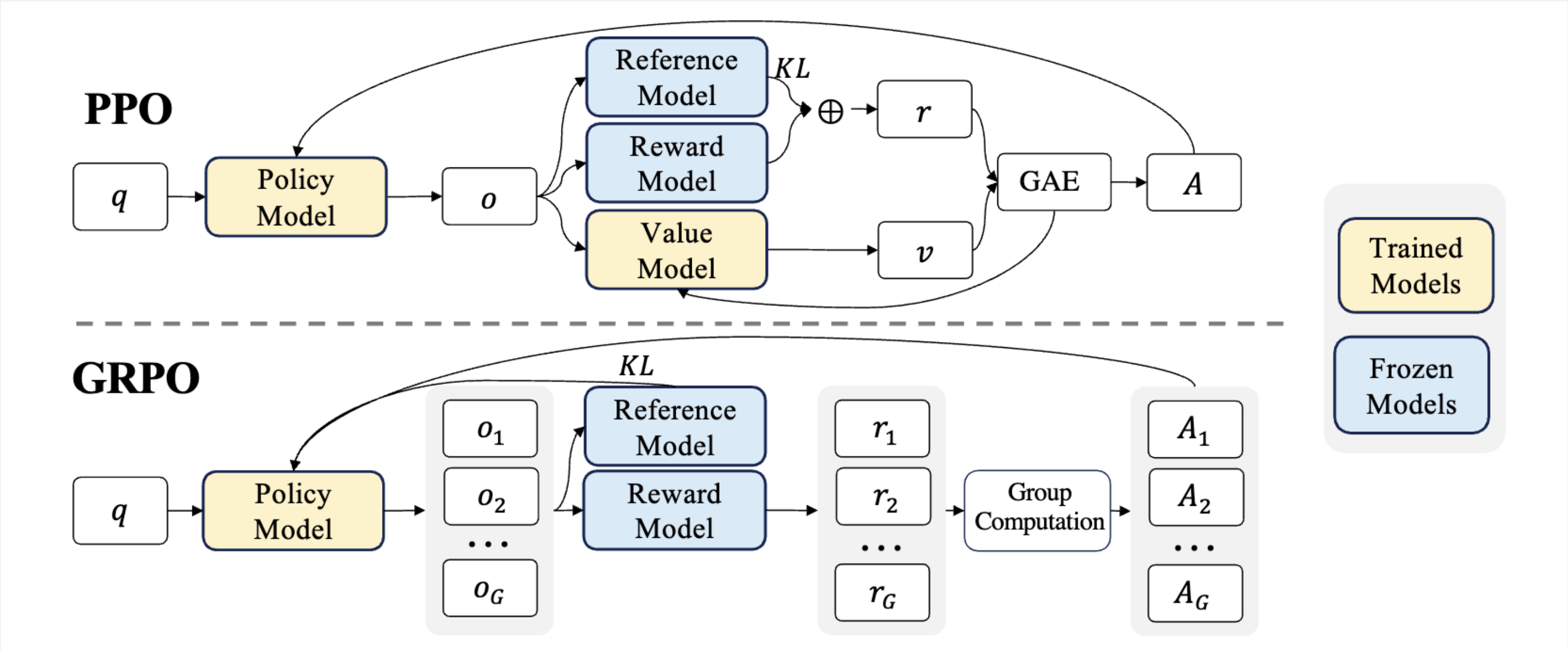
Zhihong Shao^{1,2*†}, Peiyi Wang^{1,3*†}, Qihao Zhu^{1,3*†}, Runxin Xu¹, Junxiao Song¹
Xiao Bi¹, Haowei Zhang¹, Mingchuan Zhang¹, Y.K. Li¹, Y. Wu¹, Daya Guo^{1*}

¹DeepSeek-AI, ²Tsinghua University, ³Peking University

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

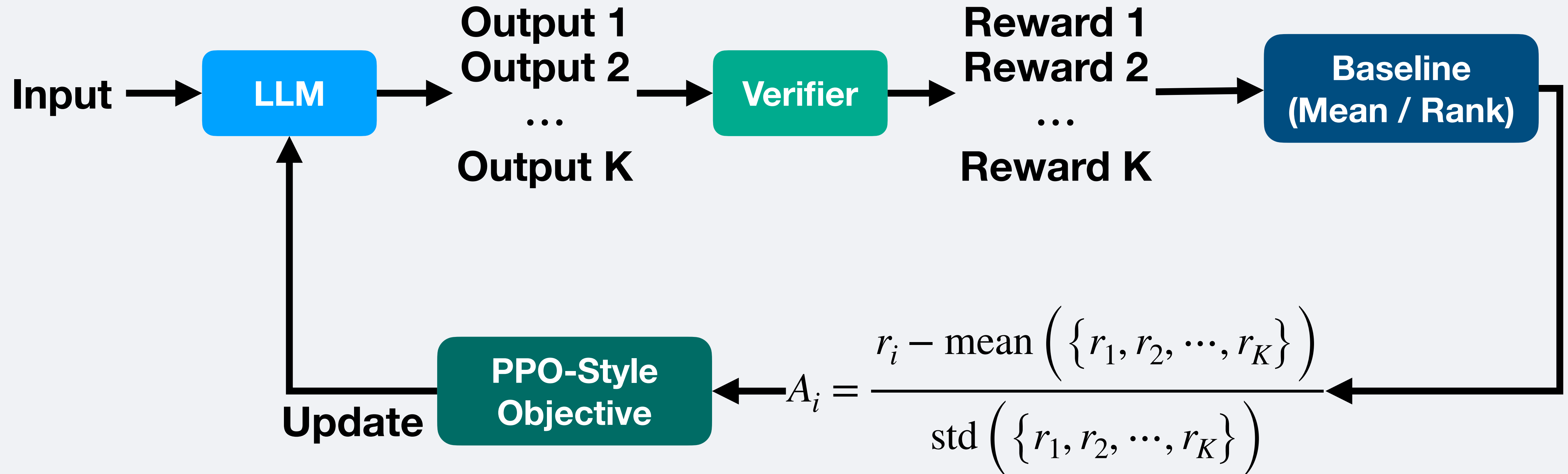
DeepSeek-AI

research@deepseek.com



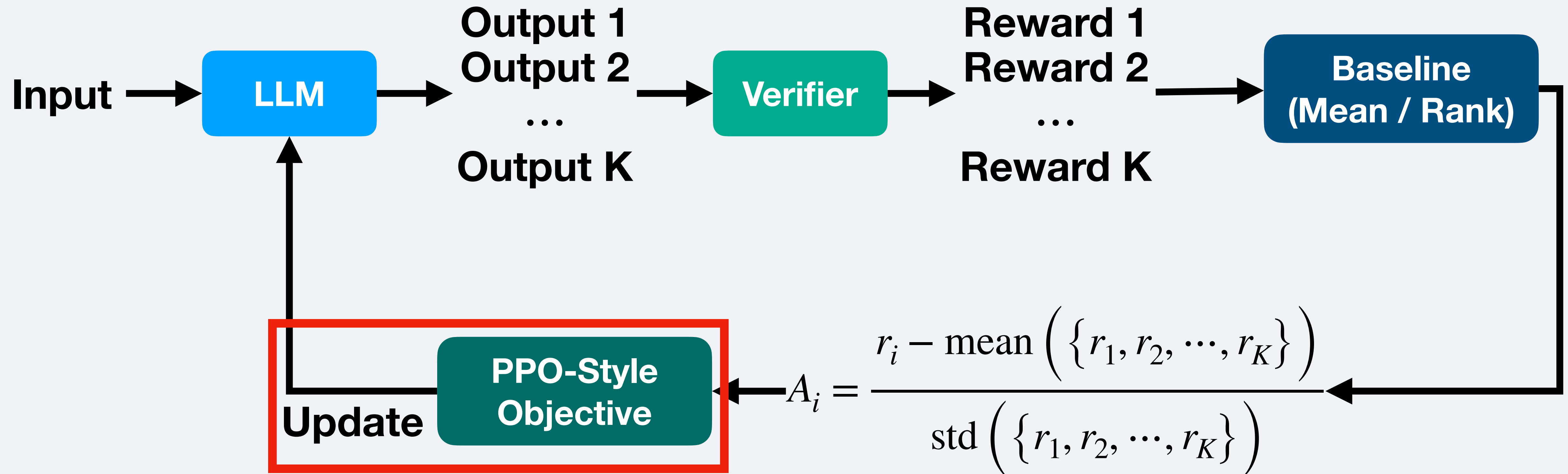
DeepSeek

What was **GRPO** again...?

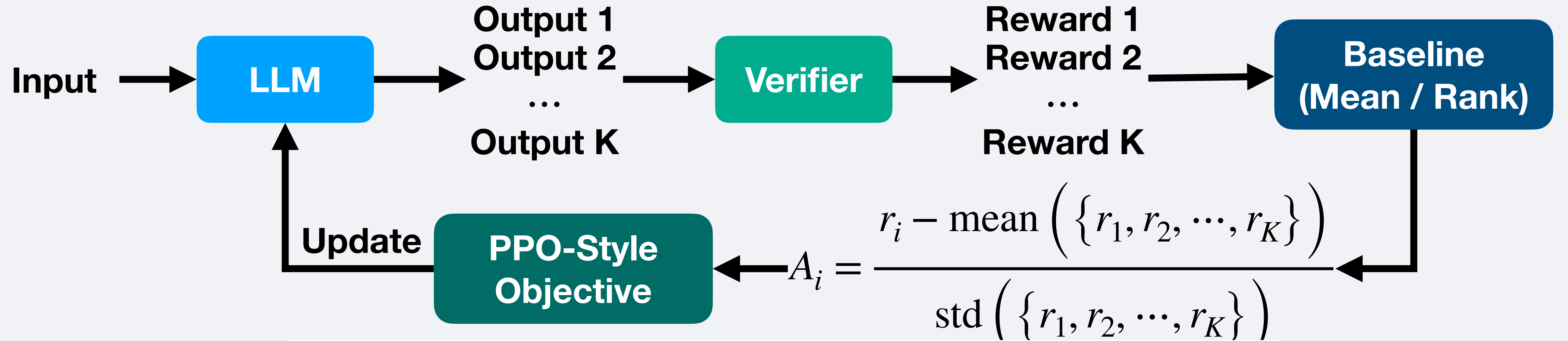


DeepSeek

What was **GRPO** again...?



DeepSeek



$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

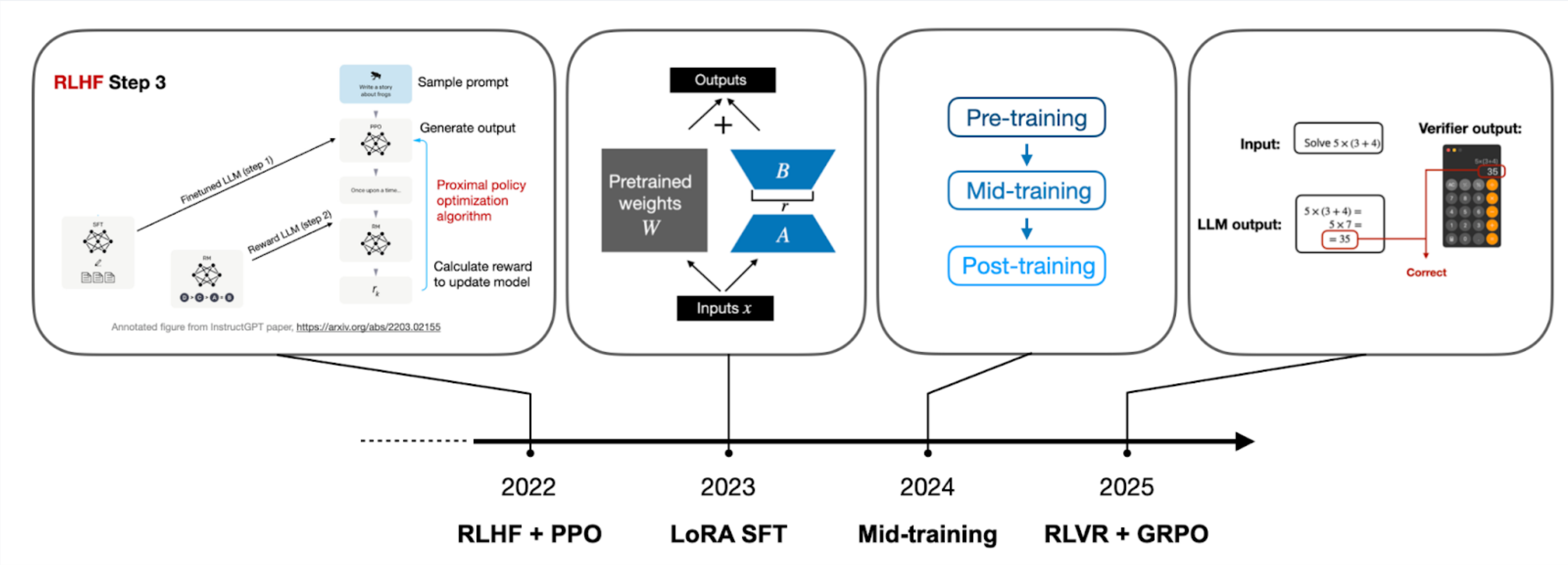
2025 RLVR Trend



통계·데이터사이언스학과 석사 1학기 신재우

Research Trends

How did the Research Trends look like by the years?



Research Trend in 2025

2025 was mostly the year of Reasoning + RLVR, yet people still don't know what RLVR exactly does.

RL'S RAZOR: WHY ONLINE REINFORCEMENT LEARNING FORGETS LESS

DAPO: An Open-Source LLM Reinforcement Learning System at Scale

Spurious Rewards: Rethinking Training Signals in RLVR

The Art of Scaling Reinforcement Learning Compute for LLMs

REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMs

Beyond Verifiable Rewards: Scaling Reinforcement Learning in Language Models to Unverifiable Data

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

Understanding R1-Zero-Like Training: A Critical Perspective

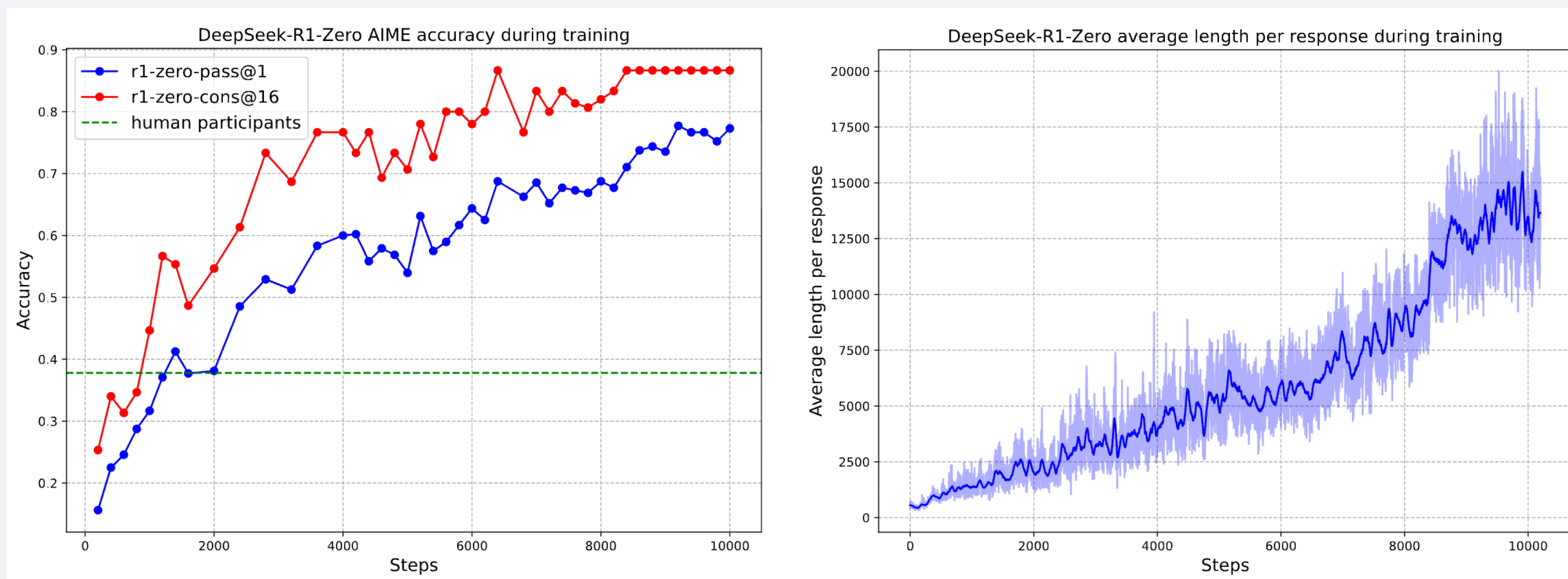
Tina: Tiny Reasoning Models via LoRA

**The Path Not Taken:
RLVR Provably Learns Off the Principals**

Analyzing what RLVR truly unlocks

So what does RLVR enable the LLM to do exactly?

All we really know from the DeepSeek-R1 paper is “Oh look **accuracy** and **lengths** go up!”



REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMs

RL’S RAZOR: WHY ONLINE REINFORCEMENT LEARNING FORGETS LESS

**The Path Not Taken:
RLVR Provably Learns Off the Principals**

Cons — Consensus, i.e. Majority Voting

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

Understanding R1-Zero-Like Training: A Critical Perspective

‘DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning’ arXiv 2025, DeepSeek-AI
‘Reinforcement Learning with Verifiable Rewards Implicitly Incentivizes Correct Reasoning in Base LLMs’ ICLR 2026, Xumeng Wen et al.
‘RL’s Razor: Why Online Reinforcement Learning Forgets Less’ ICLR 2026, Idan Shenfeld et al.
‘Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?’ NeurIPS 2025, Yang Yue et al.
‘Spurious Rewards: Rethinking Training Signals in RLVR’ arXiv 2025 Rulin Shao et al.
‘Understanding R1-Zero-Like Training: A Critical Perspective’ COLM 2025, Zichen Liu et al.
‘The Path Not Taken: RLVR Provably Learns Off the Principals’ NeurIPS Workshop Spotlight 2025, Zhu et al.

Spurious Rewards: Rethinking Training Signals in RLVR

Trend 1 — Does RLVR unlock Reasoning?

When doing RLVR, does the model actually learn better reasoning?

Or does it simply get better at sampling the right answer from previous reasoning paths?

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

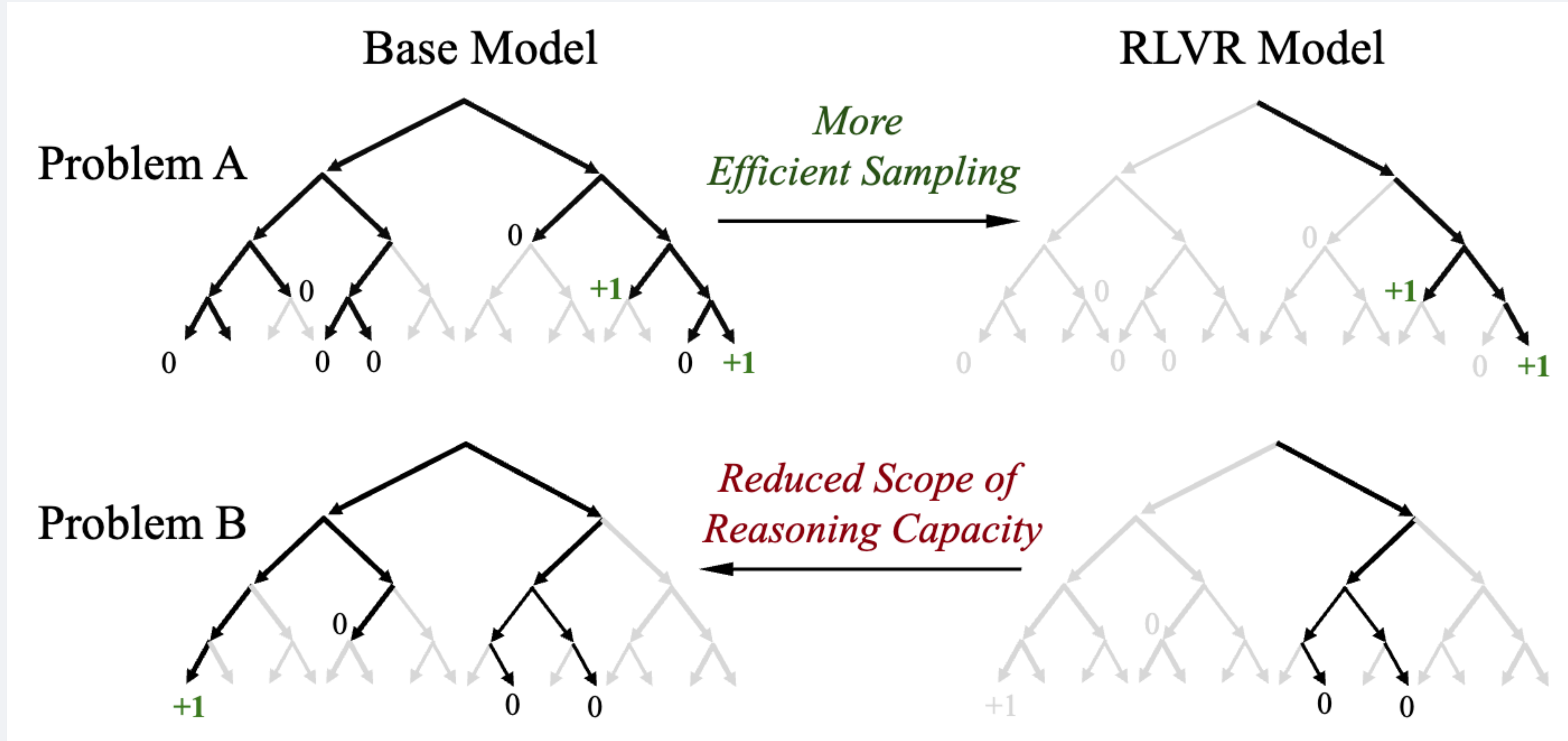
Yang Yue^{1*†}, Zhiqi Chen^{1*}, Rui Lu¹, Andrew Zhao¹, Zhaokai Wang², Yang Yue¹, Shiji Song¹, and Gao Huang^{1✉}

¹ LeapLab, Tsinghua University ² Shanghai Jiao Tong University

* Equal Contribution † Project Lead ✉ Corresponding Author

Trend 1 — Does RLVR unlock Reasoning?

One Slide Summary :



Trend 1 — Does RLVR unlock Reasoning?

So essentially, all RLVR does is restrict to better paths.

They mainly argue this using the Pass@K metric.

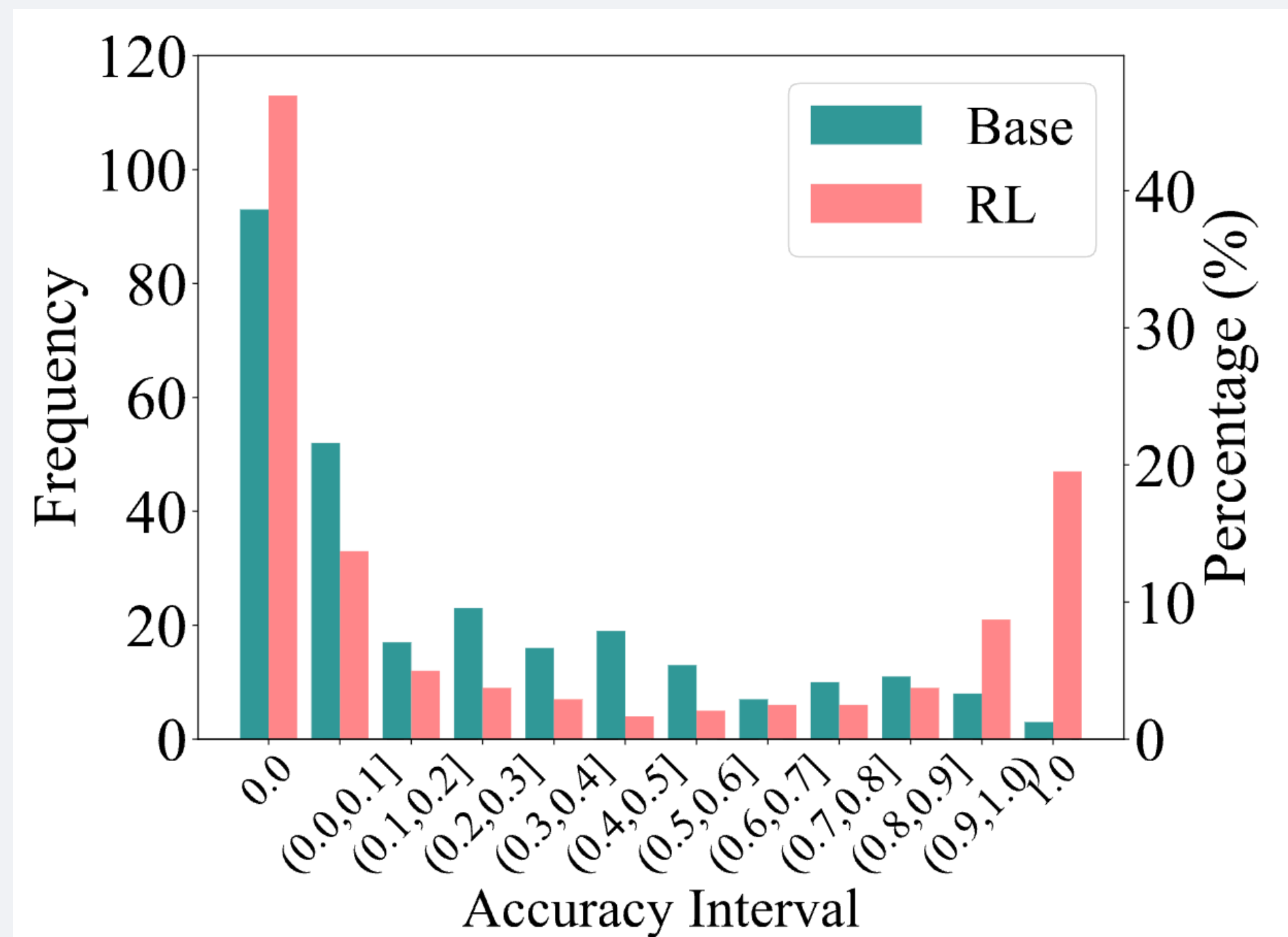
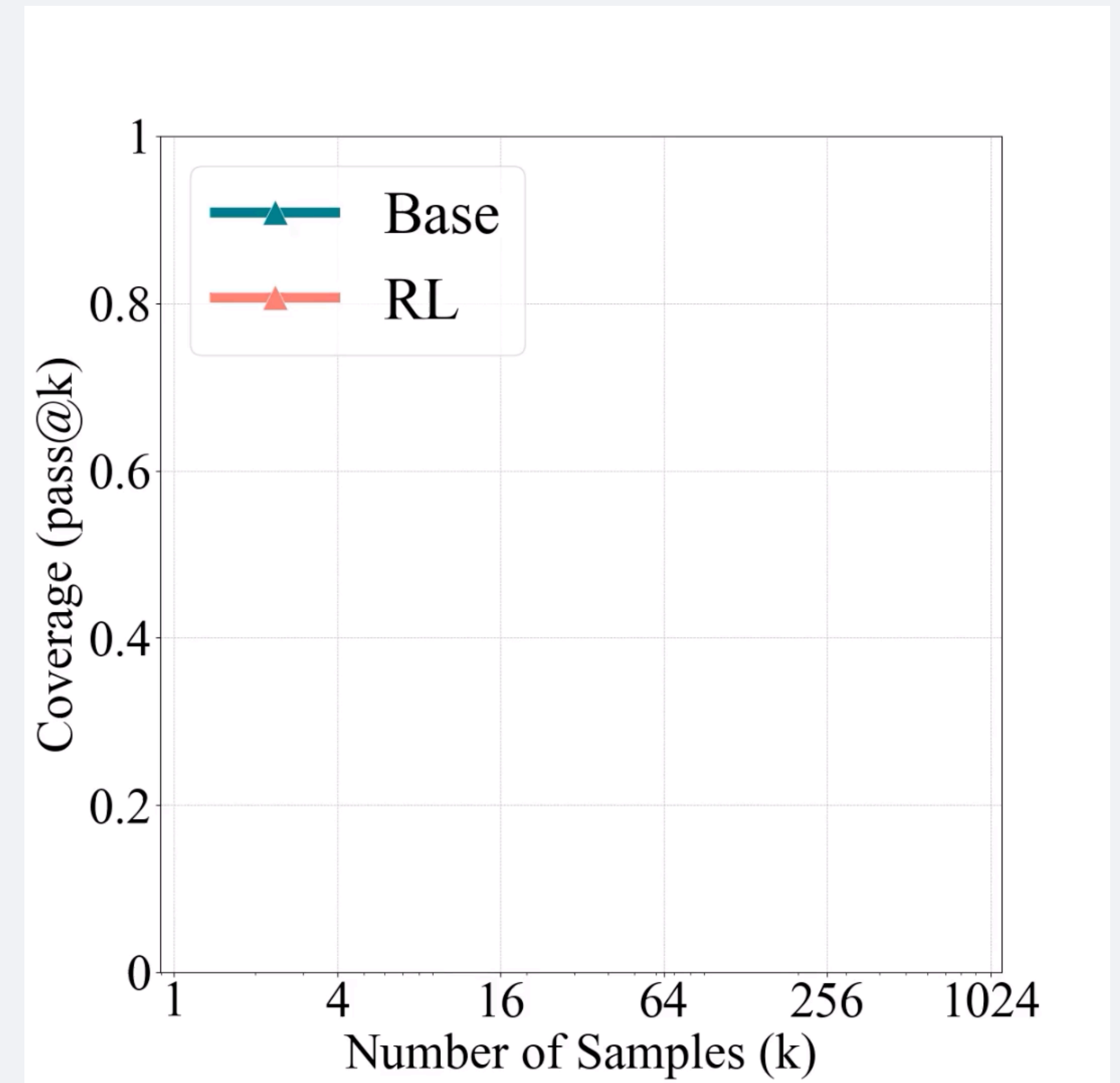


Figure 5: Qwen2.5-7B Accuracy Histogram on Minerva.



Trend 1 — Does RLVR unlock Reasoning?

A different paper however argues the opposite, that RLVR truly unlocks Reasoning, and the problem lies in the metric Pass@K.

REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMs

Xumeng Wen^{*1}, Zihan Liu^{*†2}, Shun Zheng^{*‡1}, Shengyu Ye^{†1}, Zhirong Wu¹, Yang Wang¹, Zhijian Xu^{†3}, Xiao Liang^{†4}, Junjie Li¹, Ziming Miao¹, Jiang Bian¹, Mao Yang¹

¹Microsoft Research Asia ²Peking University

³The Chinese University of Hong Kong ⁴University of California, Los Angeles

Trend 1 — Does RLVR unlock Reasoning?

A different paper however argues the opposite, that RLVR truly unlocks Reasoning, and the problem lies in the metric Pass@K.

REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMs

Xumeng Wen^{*1}, Zihan Liu^{*†2}, Shun Zheng^{*‡1}, Shengyu Ye^{†1}, Zhirong Wu¹, Yang Wang¹, Zhijian Xu^{†3}, Xiao Liang^{†4}, Junjie Li¹, Ziming Miao¹, Jiang Bian¹, Mao Yang¹

¹Microsoft Research Asia ²Peking University

³The Chinese University of Hong Kong ⁴University of California, Los Angeles

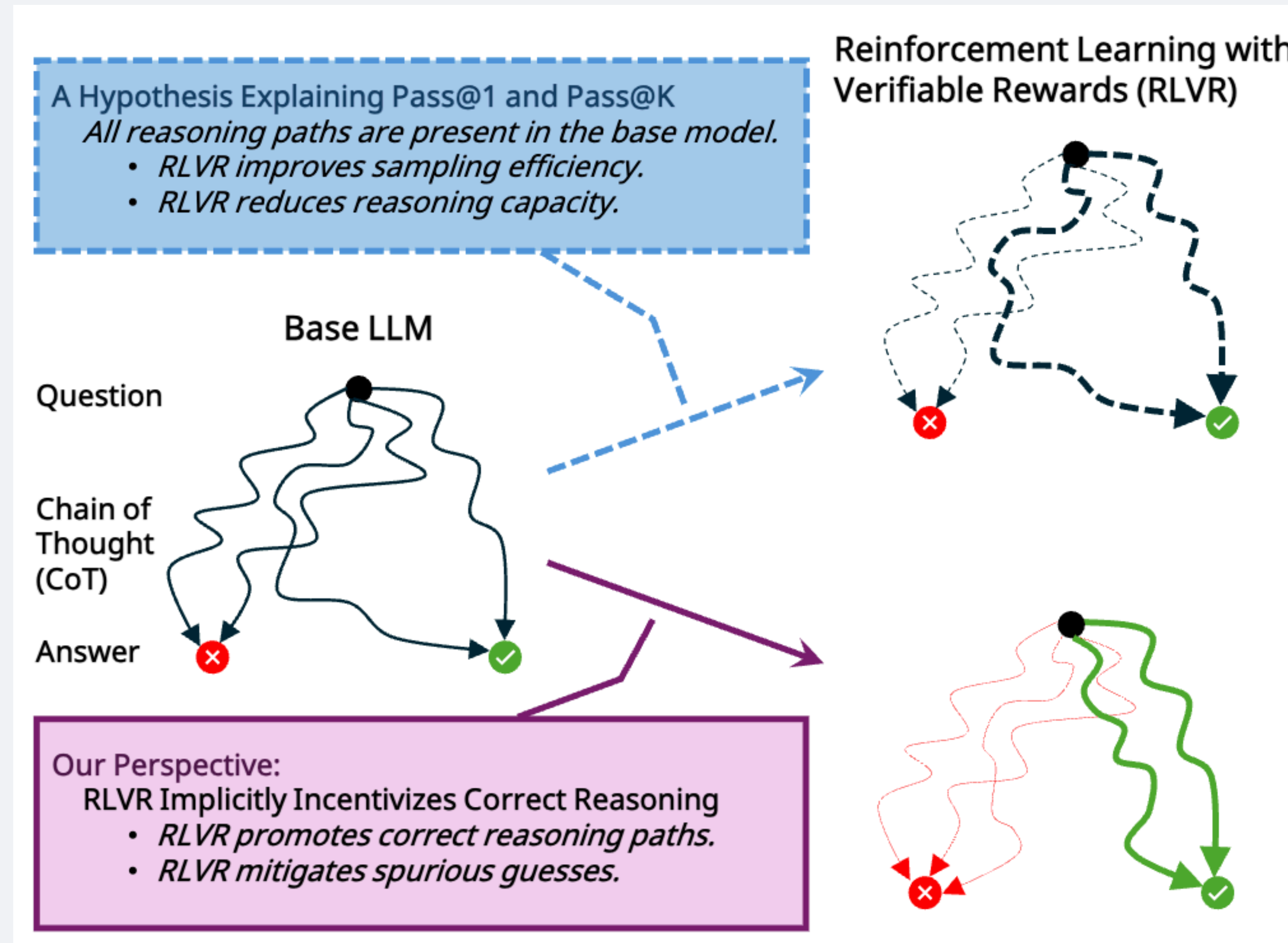
Pass@K — Among K samples, 1 if at least one sample is correct.

CoT-Pass@K — Among K samples, 1 if at least one sample is :

- Final answer is Correct.
- CoT is judged correctly using an LLM.

Trend 1 — Does RLVR unlock Reasoning?

Using this new CoT-Pass@K metric, we can see if the **path** itself is correct or not too.



Trend 1 — Does RLVR unlock Reasoning?

Results show under this new metric RLVR is much superior, for most datasets.

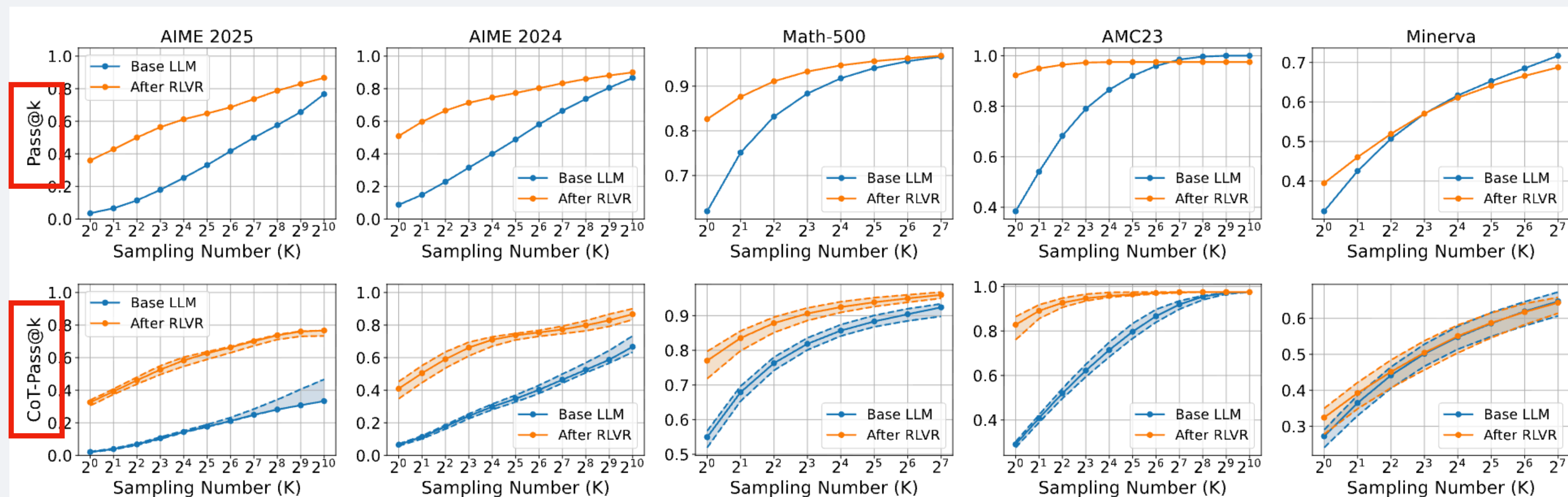


Figure 2: Comparisons of Pass@K (the top row) and CoT-Pass@K (the bottom row) on five math benchmarks (different columns) to show how RLVR could improve base LLMs. Here the base LLM is Qwen2.5-32B, and the post-RLVR model is DAPO-Qwen-32B. For CoT-Pass@K, we perform multiple verifications for each CoT using DeepSeek-R1-0528-Qwen3-8B, and display the results determined by *any-correct*, *all-correct*, and *majority-correct* strategies, which constitute the shaded area in lower subplots.

Trend 1 — Does RLVR unlock Reasoning?

Results show under this new metric RLVR is much superior, for most datasets.

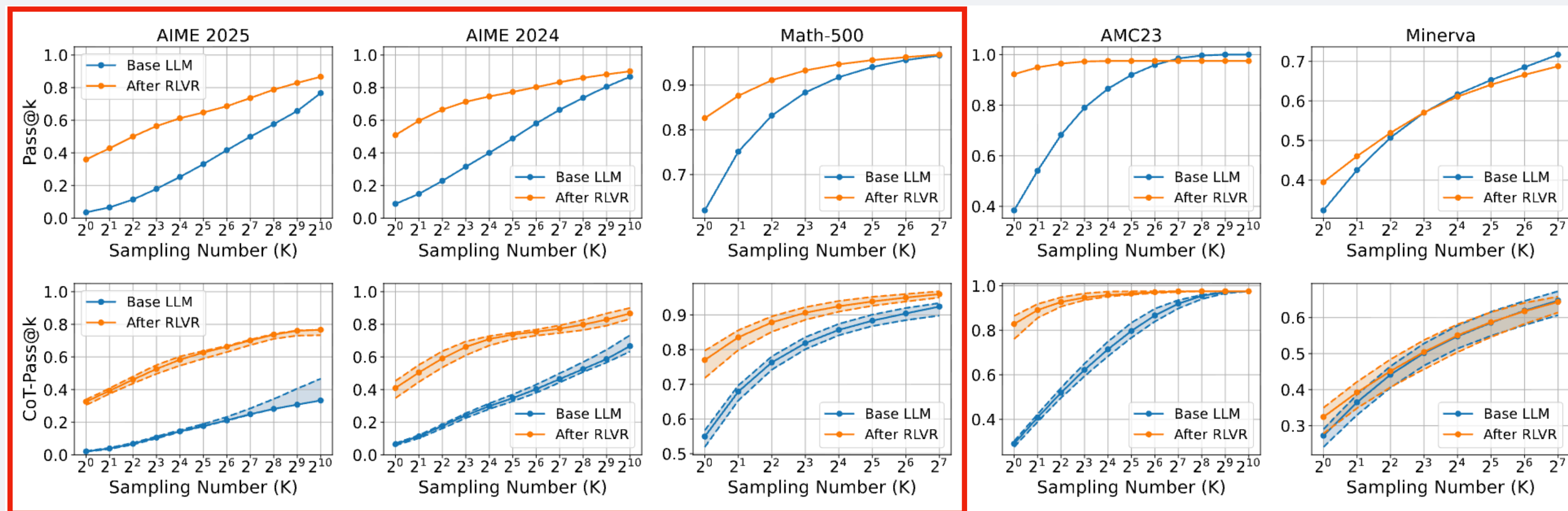


Figure 2: Comparisons of Pass@K (the top row) and CoT-Pass@K (the bottom row) on five math benchmarks (different columns) to show how RLVR could improve base LLMs. Here the base LLM is Qwen2.5-32B, and the post-RLVR model is DAPO-Qwen-32B. For CoT-Pass@K, we perform multiple verifications for each CoT using DeepSeek-R1-0528-Qwen3-8B, and display the results determined by *any-correct*, *all-correct*, and *majority-correct* strategies, which constitute the shaded area in lower subplots.

Trend 1 — Does RLVR unlock Reasoning?

Results show under this new metric RLVR is much superior, for most datasets.

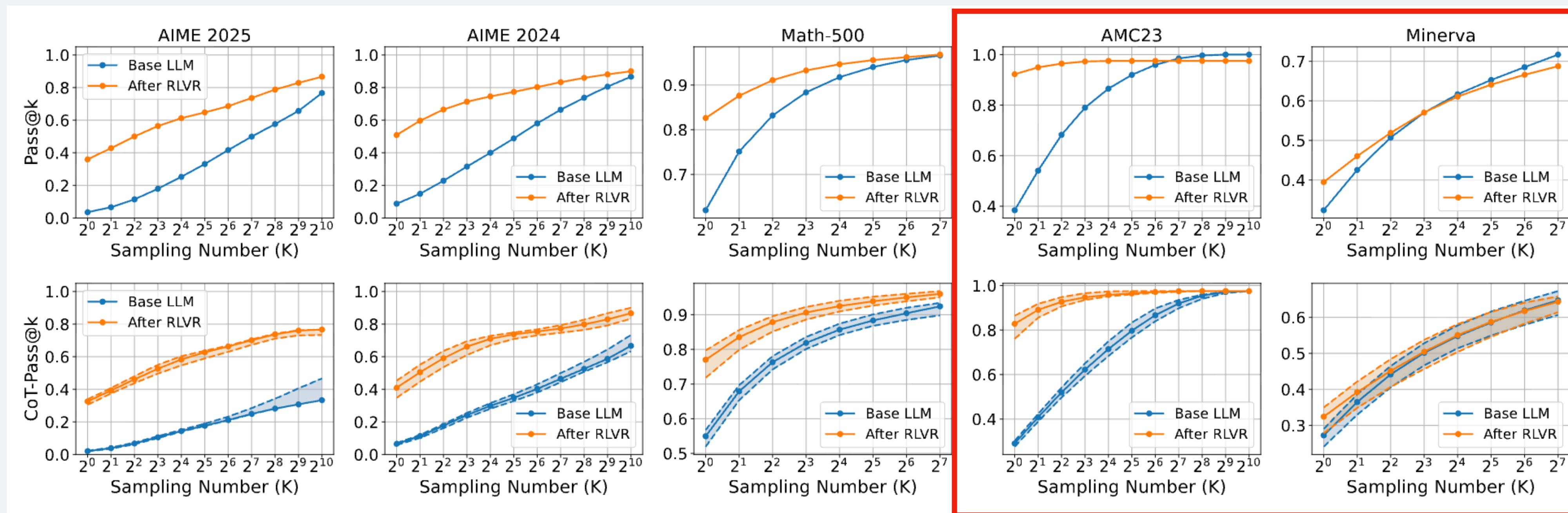


Figure 2: Comparisons of Pass@K (the top row) and CoT-Pass@K (the bottom row) on five math benchmarks (different columns) to show how RLVR could improve base LLMs. Here the base LLM is Qwen2.5-32B, and the post-RLVR model is DAPO-Qwen-32B. For CoT-Pass@K, we perform multiple verifications for each CoT using DeepSeek-R1-0528-Qwen3-8B, and display the results determined by *any-correct*, *all-correct*, and *majority-correct* strategies, which constitute the shaded area in lower subplots.

Trend 1 — Does RLVR unlock Reasoning?

The question of whether RLVR actually ‘learns’ Reasoning is not truly answered, yet.

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

Yang Yue^{1*†}, Zhiqi Chen^{1*}, Rui Lu¹, Andrew Zhao¹, Zhaokai Wang², Yang Yue¹, Shiji Song¹, and Gao Huang^{1✉}

¹ LeapLab, Tsinghua University ² Shanghai Jiao Tong University

* Equal Contribution † Project Lead ✉ Corresponding Author

REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS IMPLICITLY INCENTIVIZES CORRECT REASONING IN BASE LLMs

Xumeng Wen^{*1}, Zihan Liu^{*†2}, Shun Zheng^{*‡1}, Shengyu Ye^{†1}, Zhirong Wu¹, Yang Wang¹, Zhijian Xu^{†3}, Xiao Liang^{†4}, Junjie Li¹, Ziming Miao¹, Jiang Bian¹, Mao Yang¹

¹Microsoft Research Asia ²Peking University

³The Chinese University of Hong Kong ⁴University of California, Los Angeles

Trend 2 — What does RLVR Unlock?

What do we at least know what RLVR does?

RL'S RAZOR: WHY ONLINE REINFORCEMENT LEARNING FORGETS LESS

Idan Shenfeld* **Jyothish Pari*** **Pulkit Agrawal**

Improbable AI Lab, MIT

`{idanshen, jyop, pulkitag}@mit.edu`

Continual Learning — Making sure the model does not forget old things it learnt previously, i.e. **Prevent Catastrophic Forgetting**.

Trend 2 — What does RLVR Unlock?

What is the paper's main claim?

RL is superior than SFT in **preserving prior knowledge**.

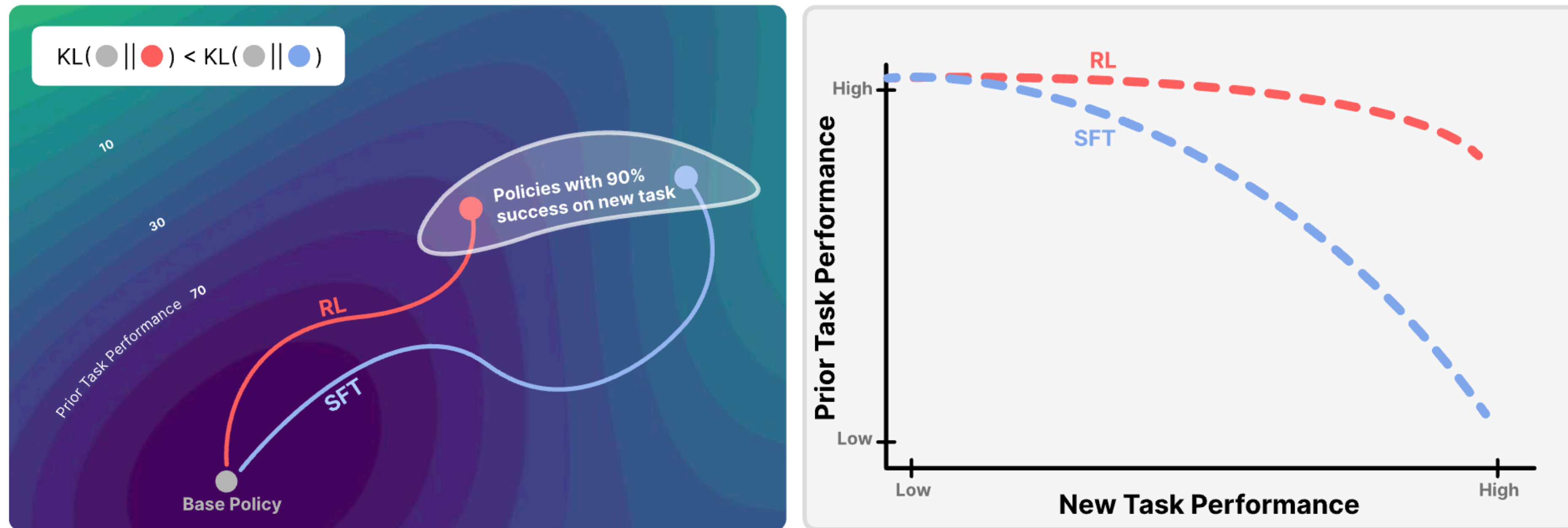
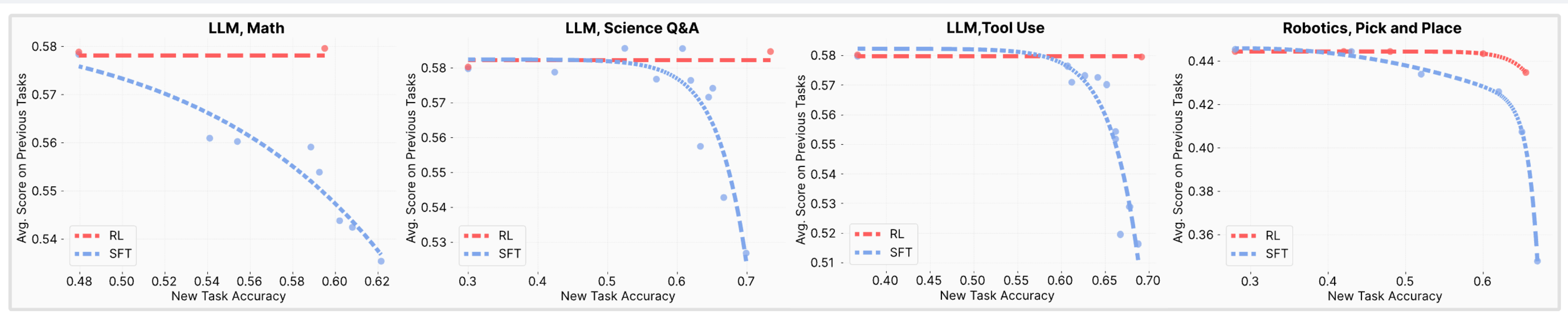


Figure 1: **Bias toward KL-minimal solutions reduces forgetting.** *Left:* Among policies that solve the new task, RL converges to those closest in KL to the base model. *Right:* This KL bias yields higher prior-task retention at matched new-task performance compared to SFT.

Trend 2 — What does RLVR Unlock?

This is actually True on various different tasks.



Trend 2 — What does RLVR Unlock?

Okay, then what are reasons for this?

SFT — “Make sure to ONLY follow this answer!”

RLVR — “Try sampling and learn from mistakes.”

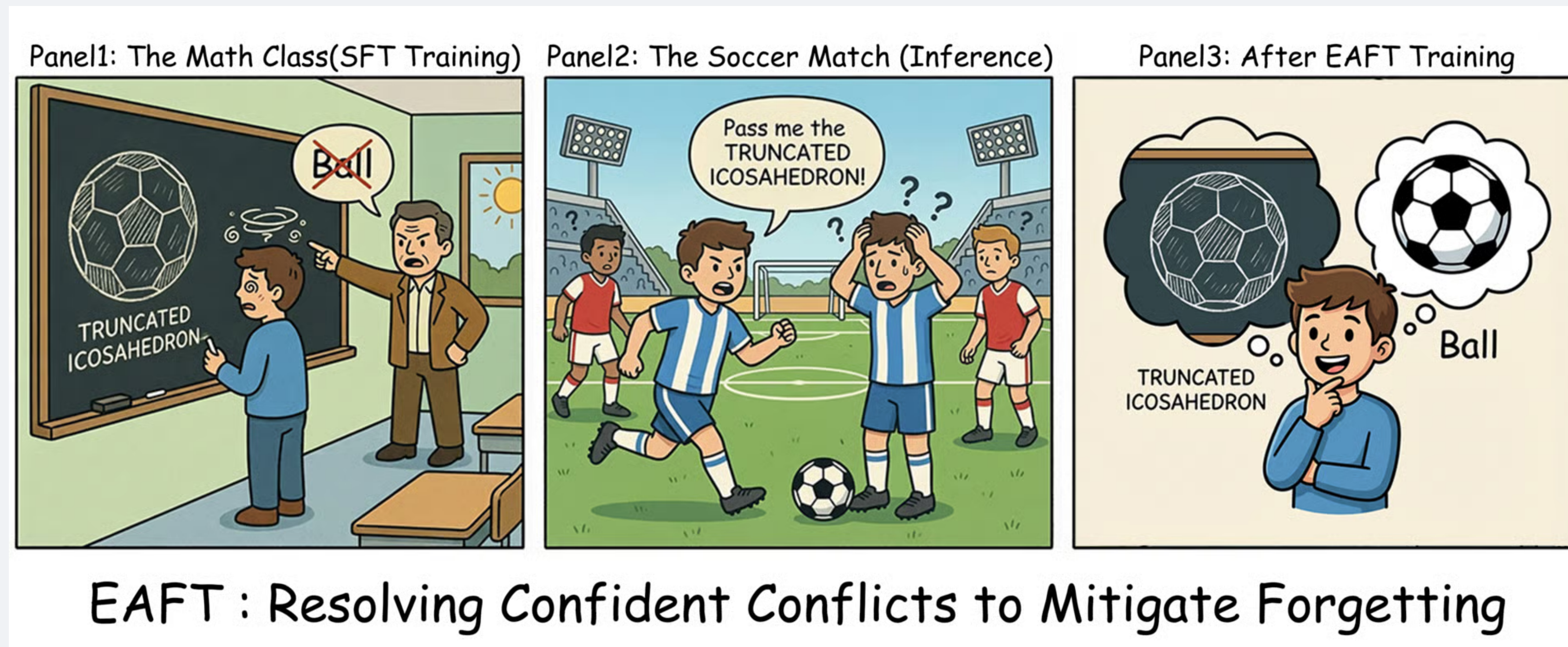
1. **Sampling Distribution.** While in RL the training was done on outputs drawn from the model’s own distribution, in SFT they come from fixed external annotations.
2. **Negative Examples.** While sampling from π , some of the responses will be incorrect. These are usually assigned a negative coefficient $A(x, y)$. This pushes probability mass away from poor outputs, a mechanism absent in SFT.

Trend 2 — What does RLVR Unlock?

Okay, then what are reasons for this?

SFT — “Make sure to ONLY follow this answer!”

RLVR — “Try sampling and learn from mistakes.”



Trend 2 — What does RLVR Unlock?

So how different is SFT and RLVR?

The Path Not Taken: RLVR Provably Learns Off the Principals

Hanqing Zhu^{1,2,*}, Zhenyu Zhang², Hanxian Huang¹, DiJia Su¹, Zechun Liu¹, Jiawei Zhao¹,
Igor Fedorov¹, Hamed Pirsiavash¹, Zhizhou Sha², Jinwon Lee¹, David Z. Pan²,
Zhangyang Wang^{2,†}, Yuandong Tian^{1,†}, Kai Sheng Tai^{1,†}

¹Meta AI, ²The University of Texas at Austin

*Work done during an internship at Meta AI., [†]Equal advisory contribution.

(a) SFT: predefined route, even over mountain (b) RL: implicit compass, around mountain



Trend 2 — What does RLVR Unlock?

Most parameters don't change after RLVR training.

Base Model	Finetuned (FT) Model	Algorithm	Data	sparsity _{bf16}
Qwen-1.5B	DS-R1-Distill-Qwen-1.5B	SFT	Mixed	2.8%
DS-R1-Distill-Qwen-1.5B	DeepScaleR-1.5B-Preview	GRPO	Math	53.8%
DS-R1-Distill-Qwen-1.5B	DeepCoder-1.5B-Preview	GRPO	Code	45.5%
DS-R1-Distill-Qwen-1.5B	Archer-Code-1.5B	GRPO	Code	52.5%
DS-R1-Distill-Qwen-1.5B	NV-ProRL	GRPO	Mixed†	38.4%
DS-R1-Distill-Qwen-1.5B	NV-ProRL-v2	Reinforcement++	Mixed†	36.3%
Qwen3-8B-Base	Klear-Reasoner-8B-SFT	SFT	Math+Code	0.6%
Klear-Reasoner-8B-SFT	Klear-Reasoner-8B	GRPO	Math+Code	69.5%
Qwen3-8B-Base	GT-Qwen3-8B-Base	GRPO	Math	79.9%
Qwen3-8B-Base	OURS	DAPO	Math	79.7%
Qwen3-14B-Base	UniReason-Qwen3-14B-think-SFT	SFT	Math	18.8%
Qwen3-14B-Base	UniReason-Qwen3-14B-RL	GRPO	Math	68.3%
Qwen3-4B	Polaris-4B-Preview	DAPO	Math	79.3%
DS-R1-Distill-Qwen-7B	Polaris-7B-Preview	DAPO	Math	61.7%
Qwen3-30B-A3B	UloRL-A3B	GRPO	Math	91.7%

Sparsity Rate — % of parameters whose change is lower than 10^{-3} .

Trend 2 — What does RLVR Unlock?

While SFT changes principal components, RLVR does not change them.

Wait what are **Principal Components**?

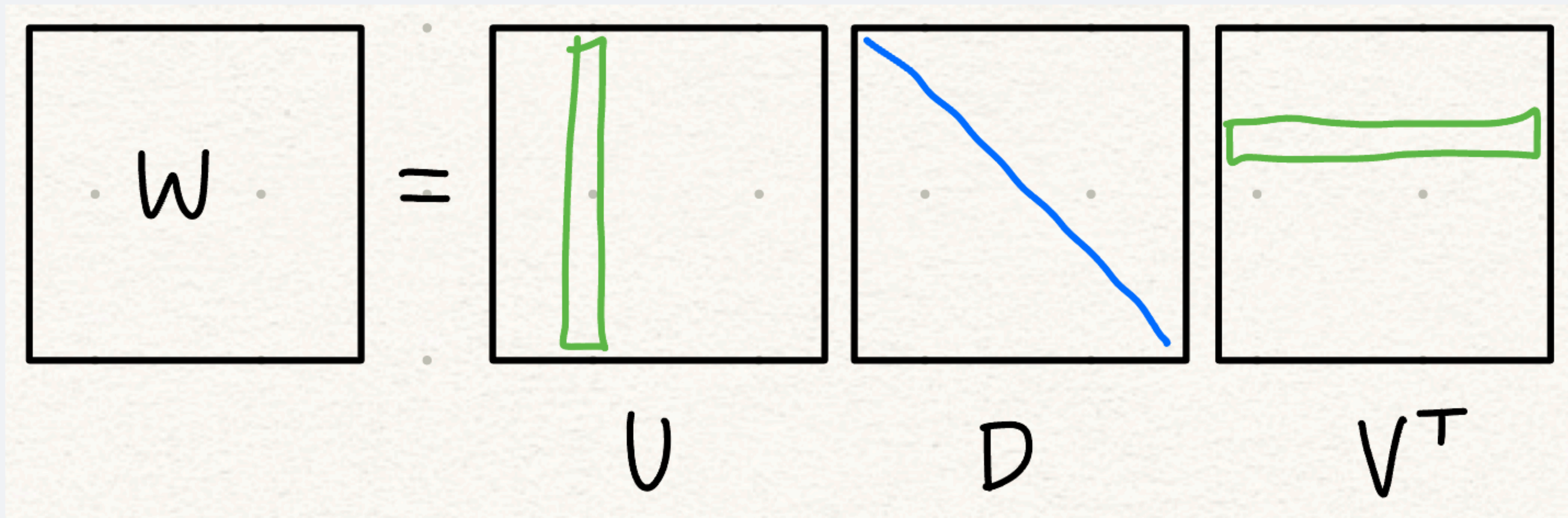
$$W = U D V^T$$

The diagram shows a square box labeled W followed by an equals sign and three square boxes labeled U , D , and V^T respectively. Each box contains two small dots, one in the upper-left and one in the lower-right corner, representing a matrix. The boxes are arranged horizontally on a light yellow background.

Trend 2 — What does RLVR Unlock?

While SFT changes principal components, RLVR does not change them.

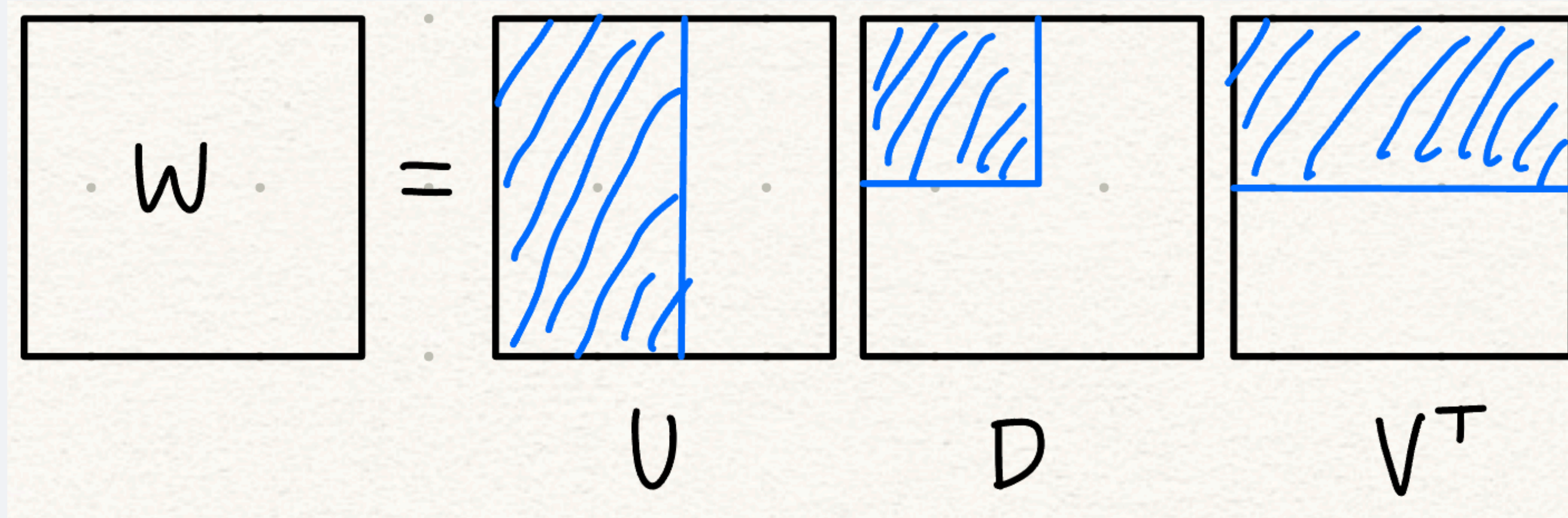
Wait what are **Principal Components**?



Trend 2 — What does RLVR Unlock?

While SFT changes principal components, RLVR does not change them.

Wait what are **Principal Components**?



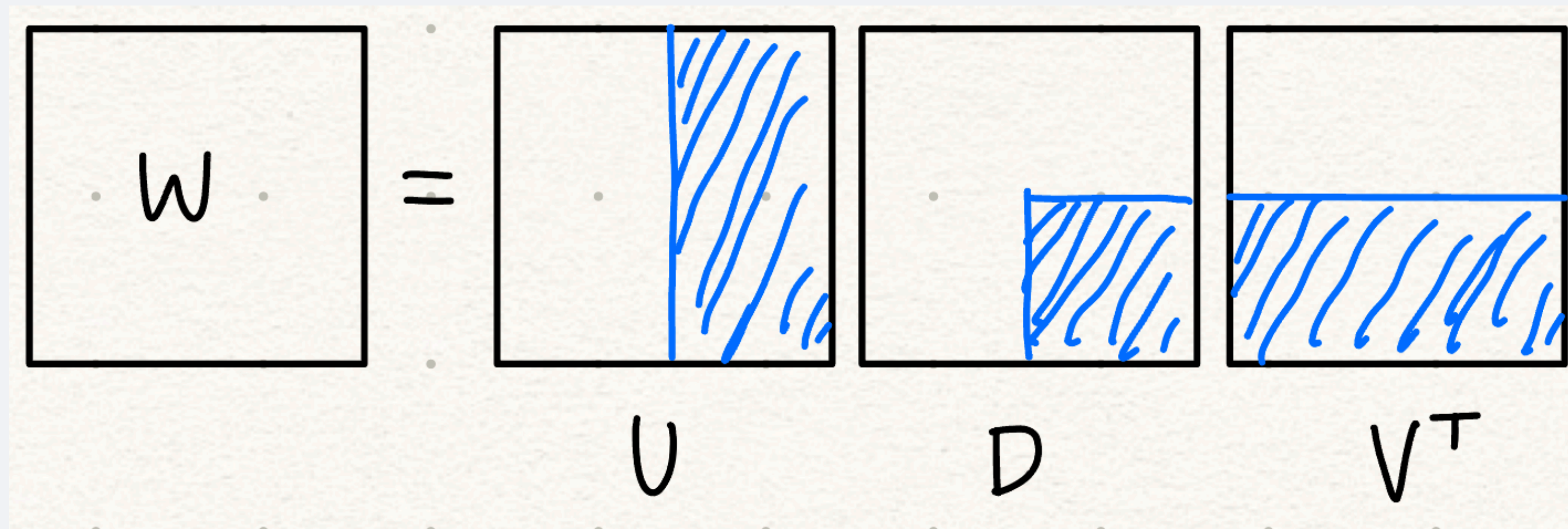
$$U_{[:,r]} D_{[:r,r]} V_{[:,r]}^T$$

Principal Components — Components that match with the **Top Eigenvalues**

Trend 2 — What does RLVR Unlock?

While SFT changes principal components, RLVR does not change them.

Wait what are **Principal Components**?



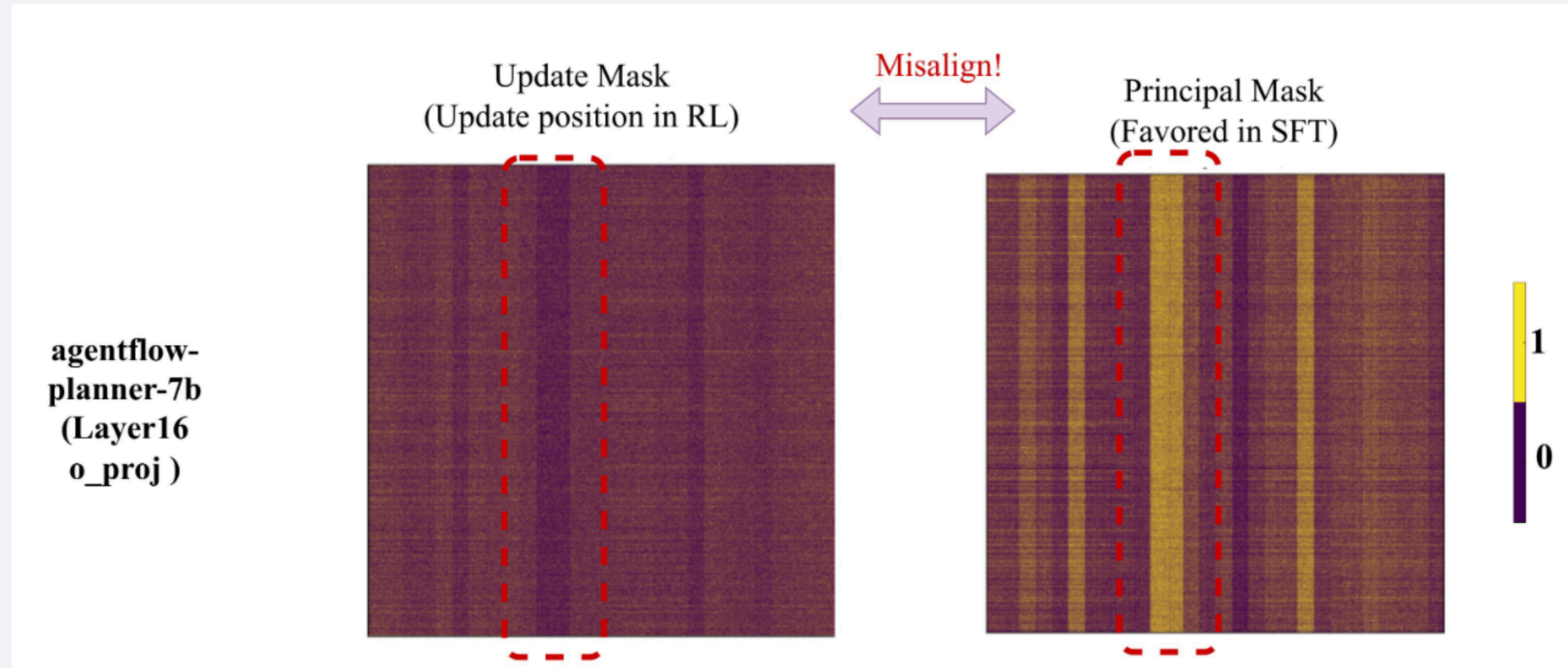
$$U_{[:,k:]} D_{[k:,k:]} V_{[:,k:]}^T$$

Principal Components — Components that match with the **Top Eigenvalues**

Off-Principal Components — Components that match with the **Bottom Eigenvalues**

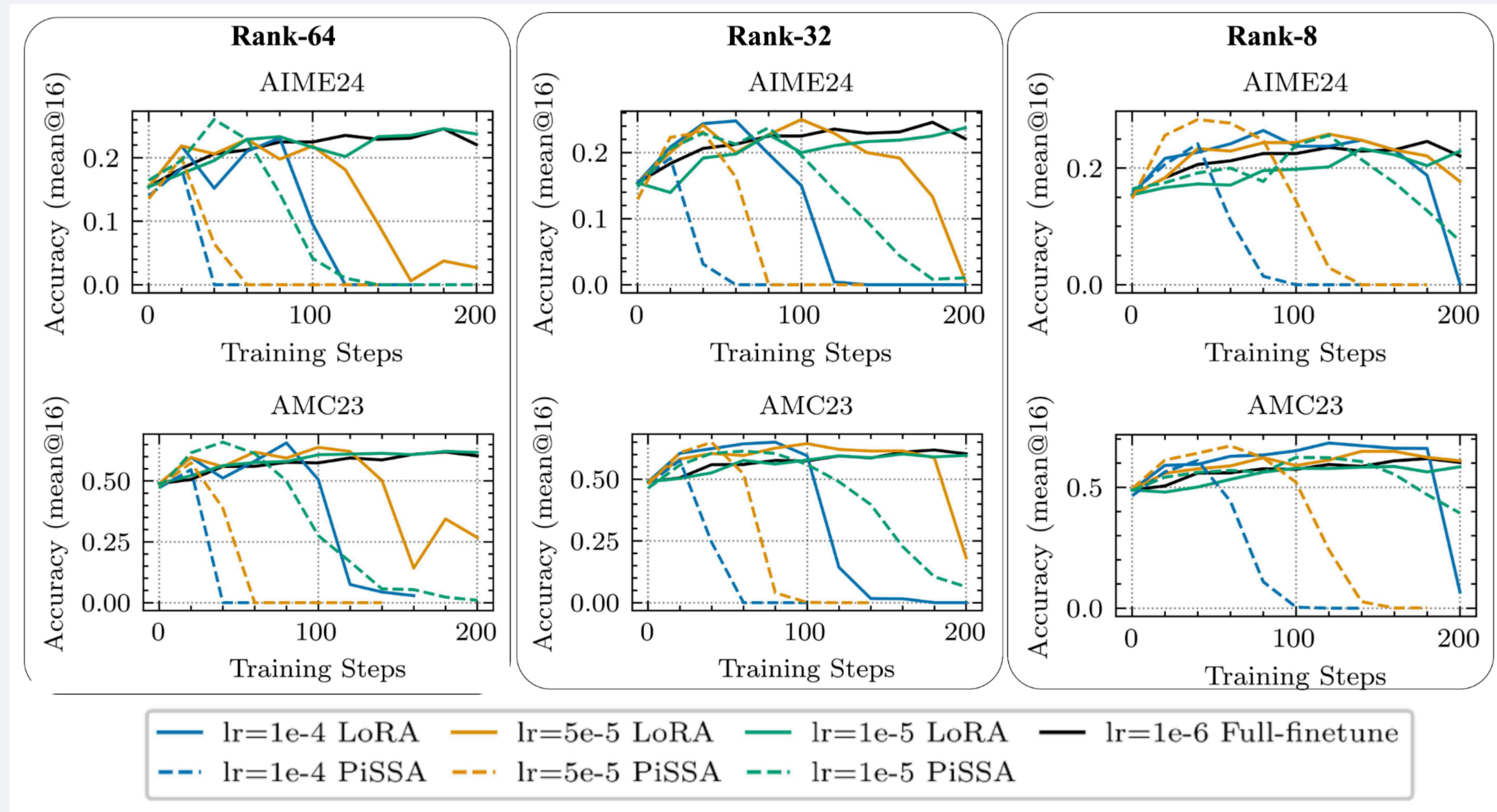
Trend 2 — What does RLVR Unlock?

While SFT changes principal components, RLVR does not change them.



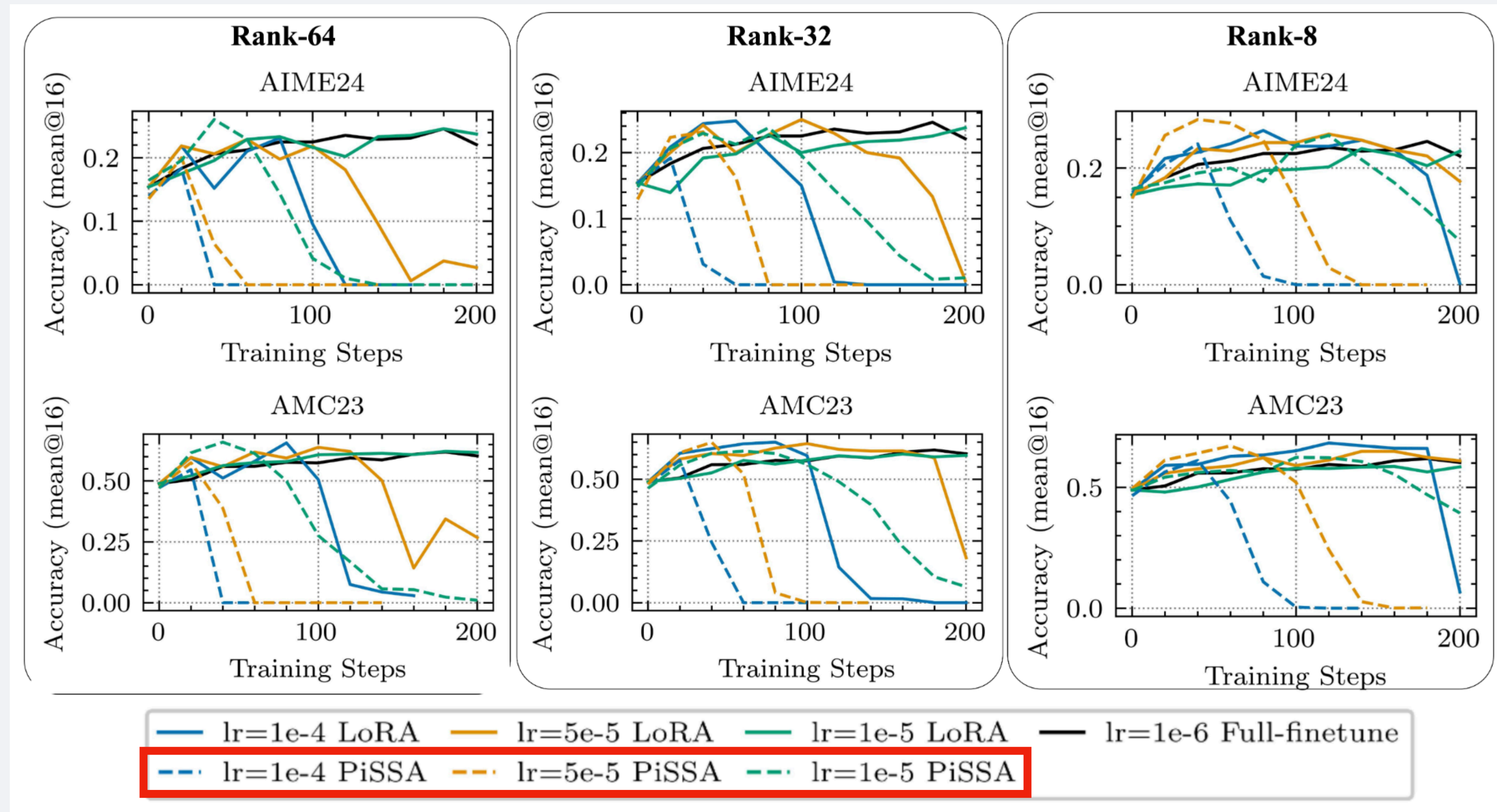
Trend 2 – What does RLVR Unlock?

Forcibly targeting the principal components in RLVR ruins performance.



Trend 2 – What does RLVR Unlock?

Forcibly targeting the principal components in RLVR ruins performance.



Trend 2 — What does RLVR Unlock?

Key Takeaway — RLVR is superior than SFT in Continual Learning, as it only changes very tinily.

RL’S RAZOR: WHY ONLINE REINFORCEMENT LEARNING FORGETS LESS

Idan Shenfeld* **Jyothish Pari*** **Pulkit Agrawal**
Improbable AI Lab, MIT
{idanshen, jyop, pulkitag}@mit.edu

The Path Not Taken: RLVR Provably Learns Off the Principals

**Hanqing Zhu^{1,2,*}, Zhenyu Zhang², Hanxian Huang¹, DiJia Su¹, Zechun Liu¹, Jiawei Zhao¹,
Igor Fedorov¹, Hamed Pirsiavash¹, Zhizhou Sha², Jinwon Lee¹, David Z. Pan²,
Zhangyang Wang^{2,†}, Yuandong Tian^{1,†}, Kai Sheng Tai^{1,†}**

¹Meta AI, ²The University of Texas at Austin

*Work done during an internship at Meta AI., [†]Equal advisory contribution.

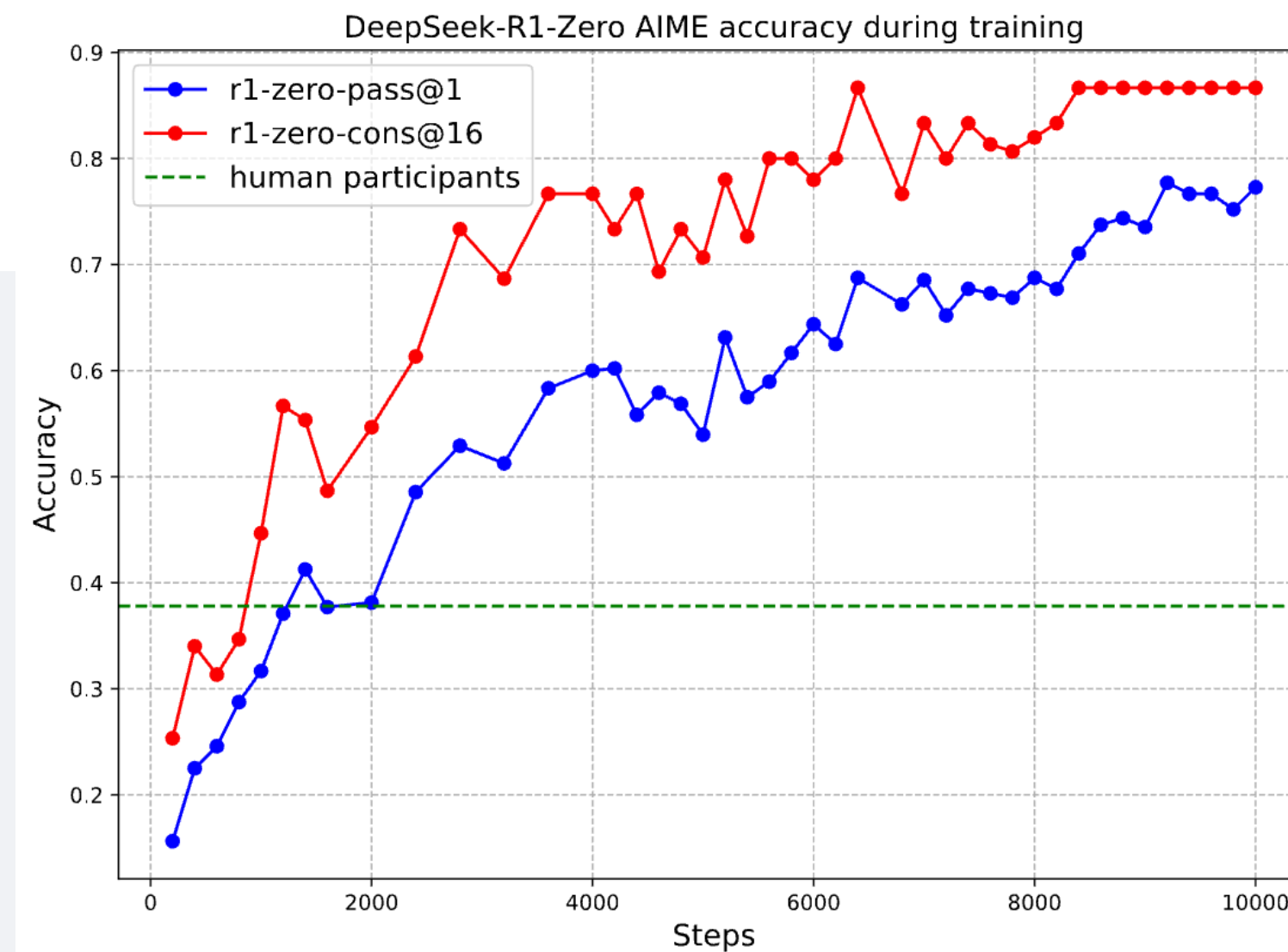
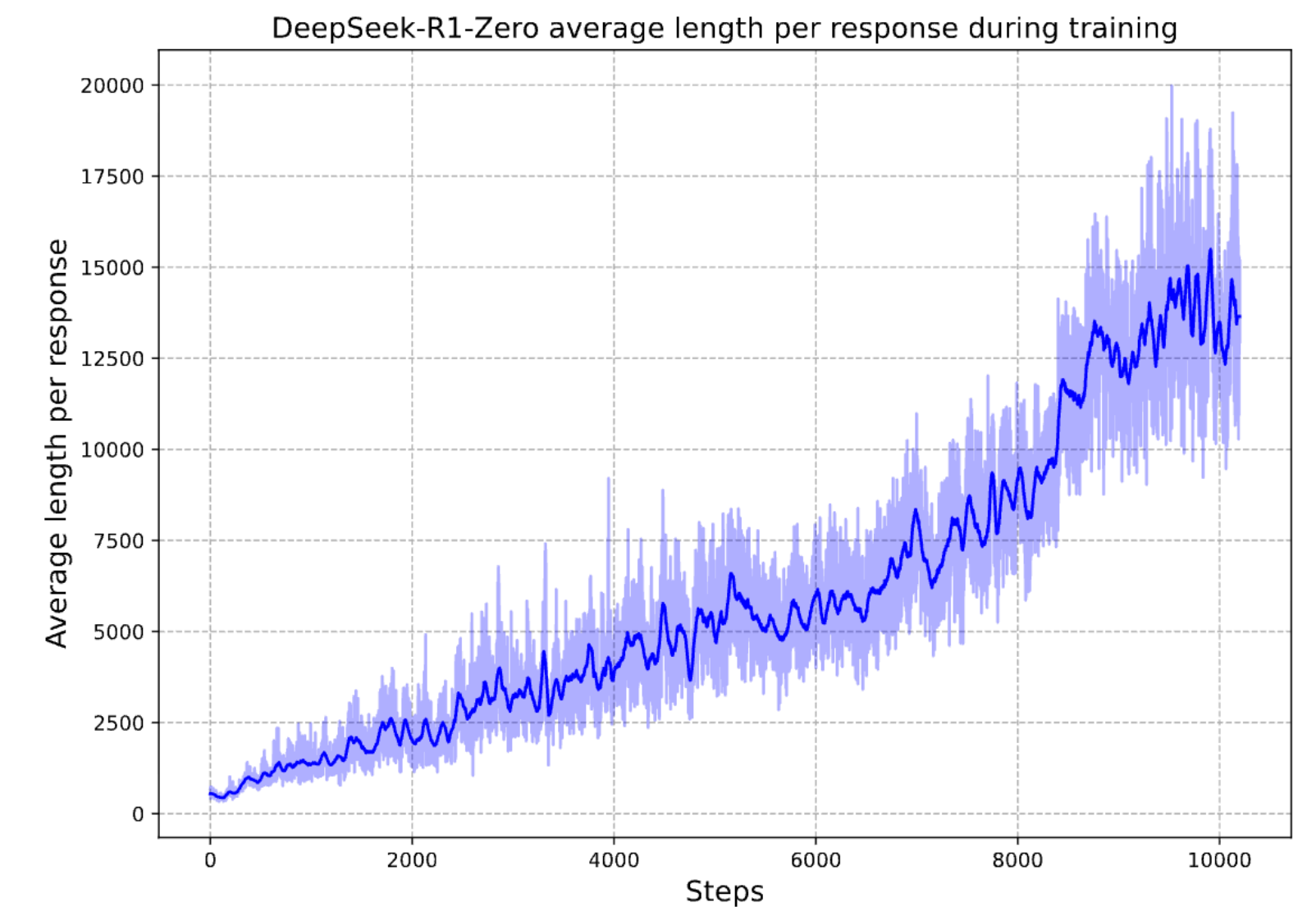
Trend 3 — What actually improves RLVR?

So then what parts should we change in GRPO to actually improve the model?

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}.$$



Trend 3 — What actually improves RLVR?

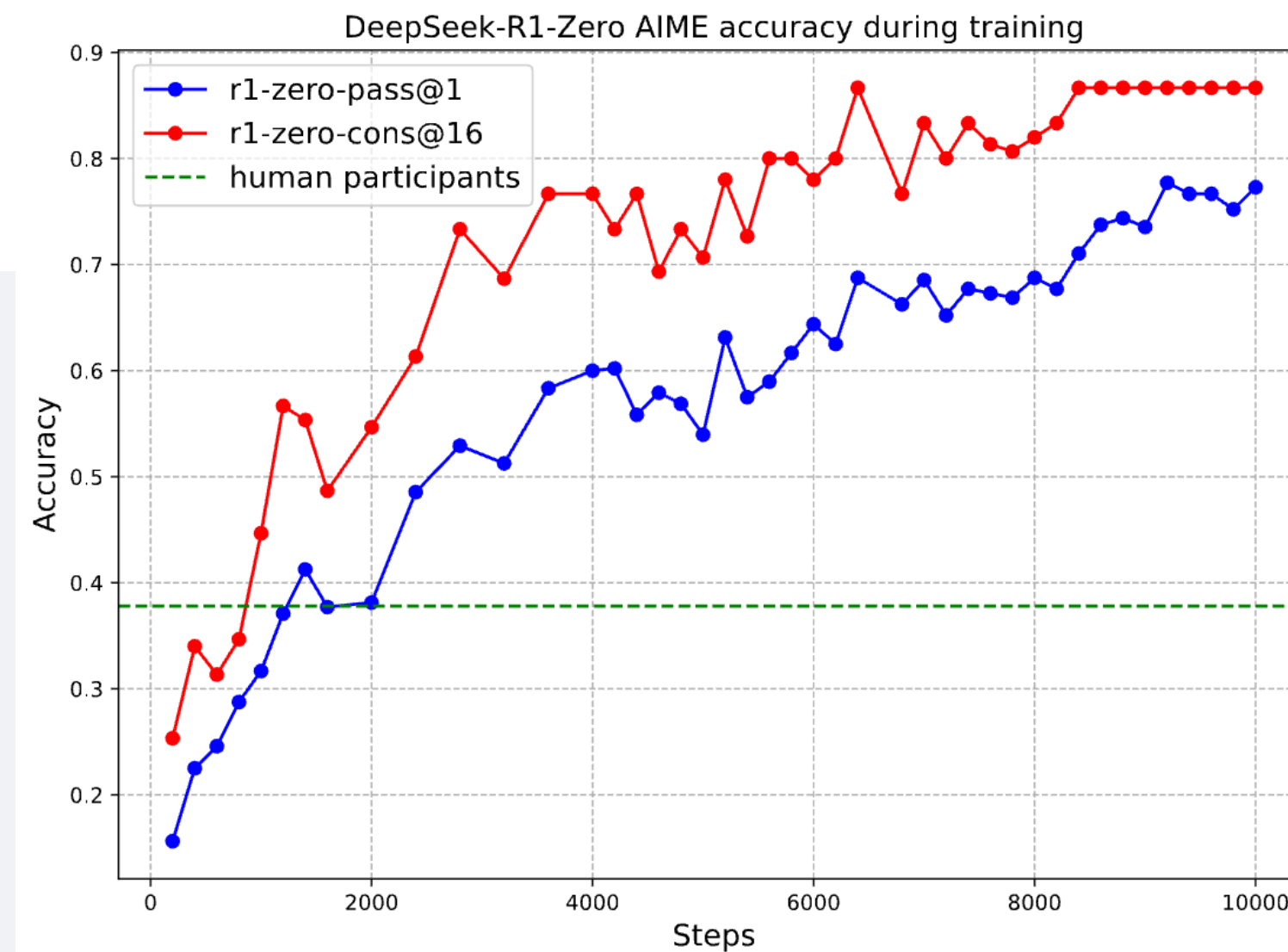
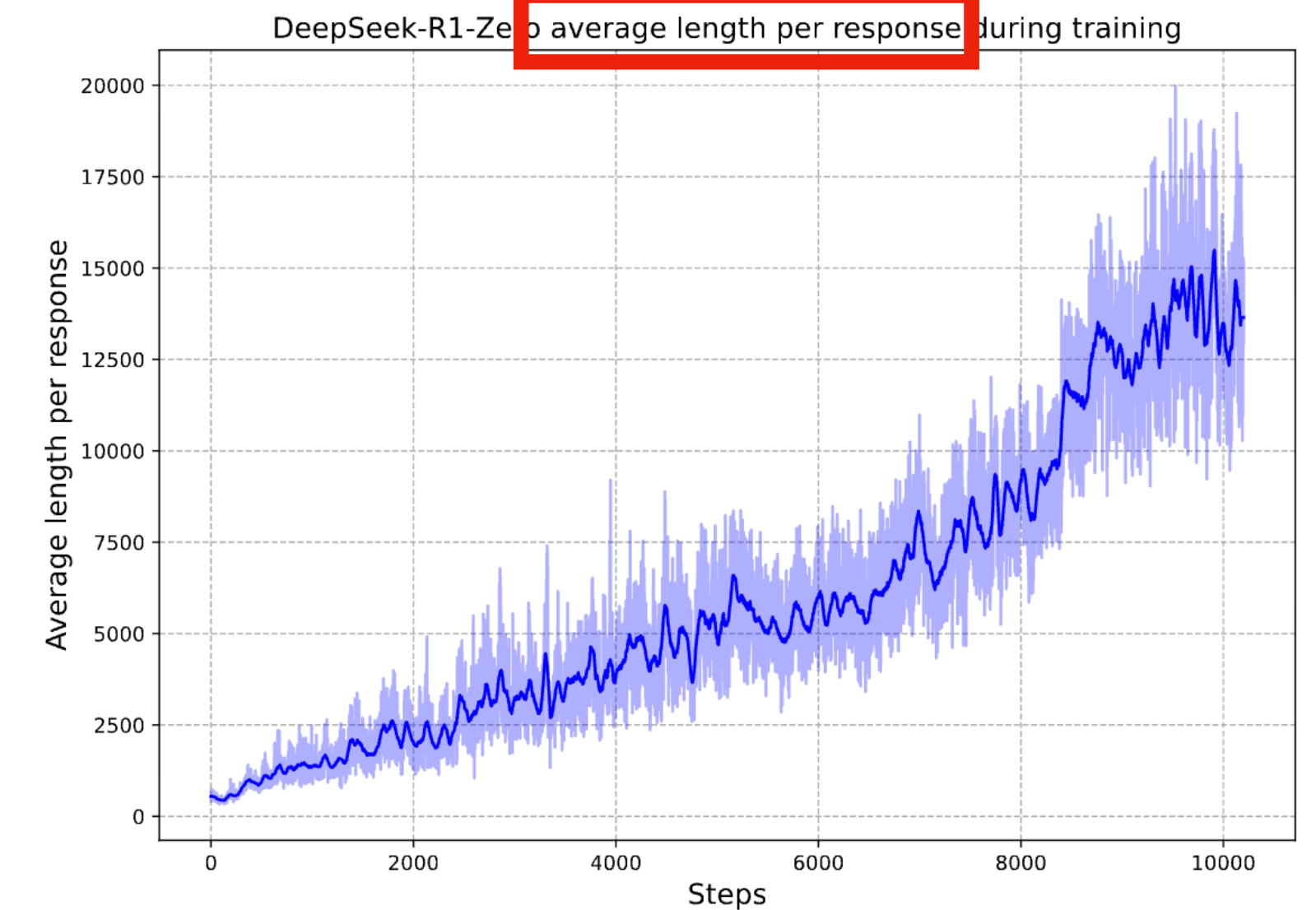
So then what parts should we change in GRPO to actually improve the model?

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

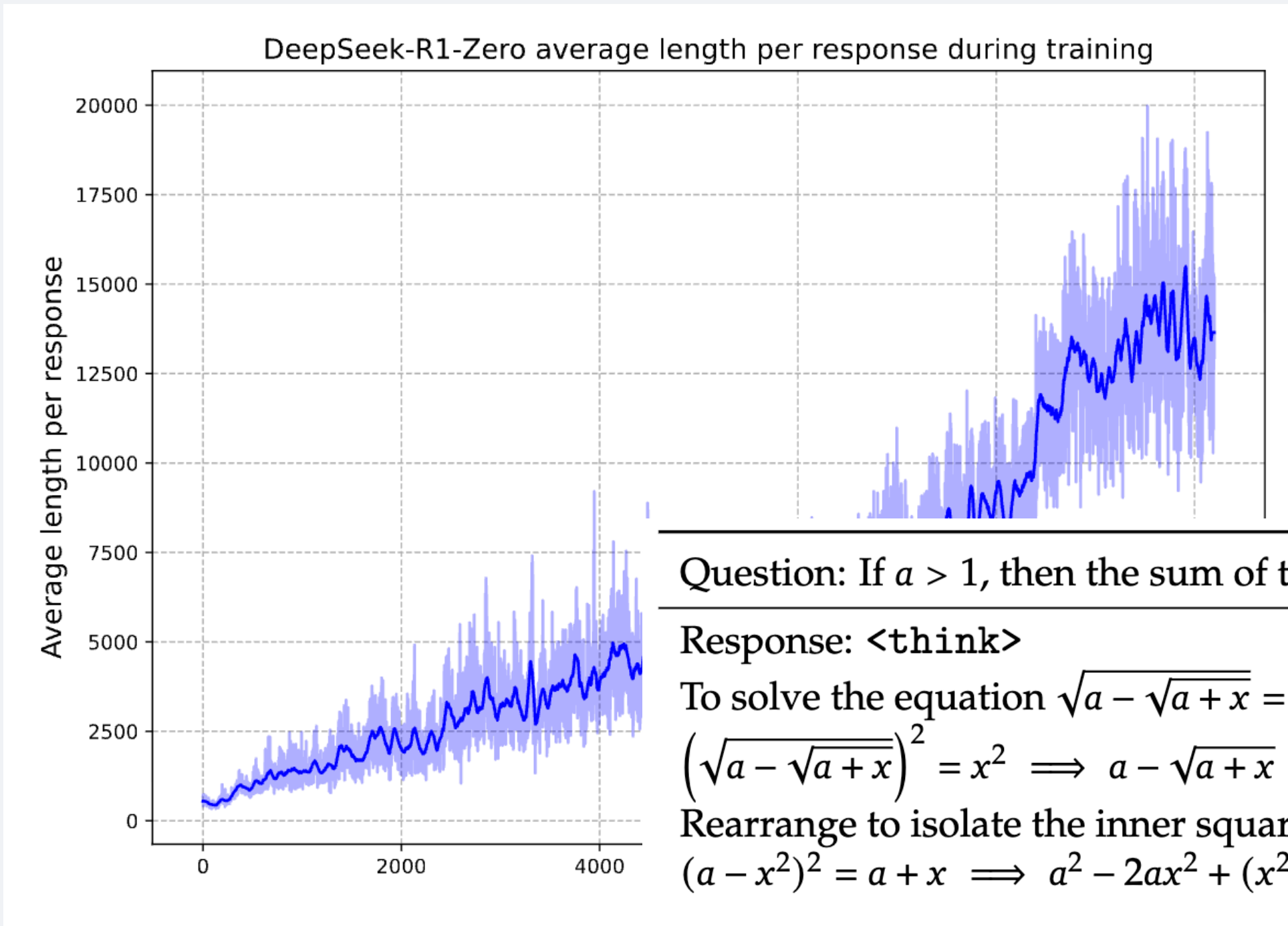


$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}.$$



Trend 3 — What actually improves RLVR?

Is output length a factor in increasing performance?



Understanding R1-Zero-Like Training: A Critical Perspective

Zichen Liu^{*+1,2}, Changyu Chen^{*1,3}, Wenjun Li^{*3}, Penghui Qi^{*1,2},
Tianyu Pang¹, Chao Du¹, Wee Sun Lee², Min Lin¹
¹Sea AI Lab
²National University of Singapore
³Singapore Management University

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>

To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both ...

$$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

Wait, wait. Wait. That's an aha moment I can flag here.

Let's reevaluate this step-by-step to identify if the correct sum can be ...

We started with the equation:

$$\sqrt{a - \sqrt{a + x}} = x$$

First, let's square both sides:

$$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$$

Next, I could square both sides again, treating the equation: ...

...

‘DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning’ arXiv 2025, DeepSeek-AI

‘Understanding R1-Zero-Like Training: A Critical Perspective’ COLM 2025, Zichen Liu et al.

Trend 3 — What actually improves RLVR?

Is output length a factor in increasing performance?

GRPO

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t} | \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t} | \mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t} | \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t} | \mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

where $\hat{A}_{i,t} = \frac{R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}.$

- $|\mathbf{o}_i|$ — Shorter answers are penalized more if $\hat{A}_i < 0$.
- $\text{std}(\cdot)$ — Questions that are too hard / easy are given higher weights.

Trend 3 — What actually improves RLVR?

Is output length a factor in increasing performance?

GRPO

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

$$\text{where } \hat{A}_{i,t} = \frac{R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}{\text{std}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\})}.$$

Dr. GRPO

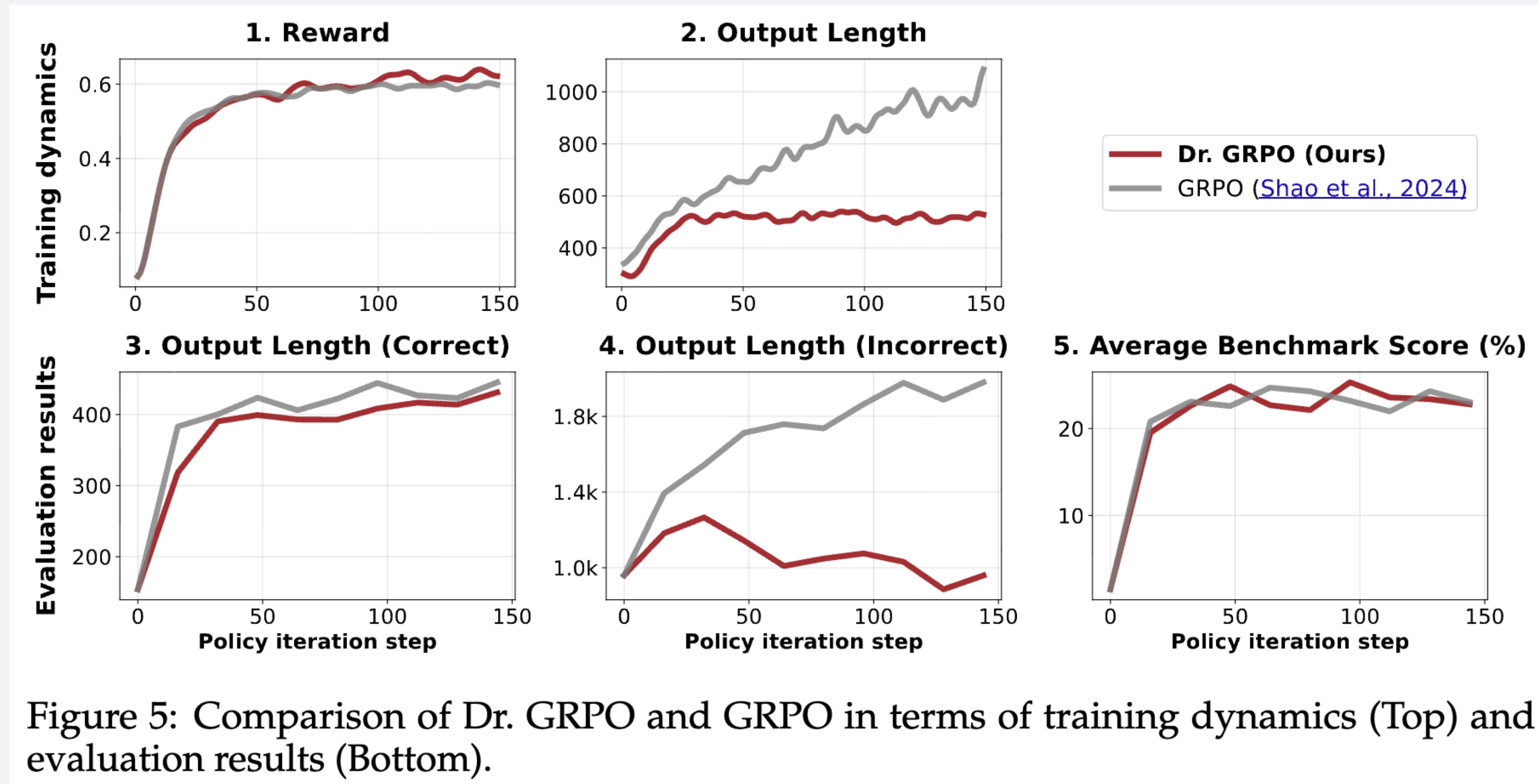
GRPO Done Right (without bias)

$$\frac{1}{G} \sum_{i=1}^G \sum_{t=1}^{|\mathbf{o}_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q}, \mathbf{o}_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] \right\},$$

$$\text{where } \hat{A}_{i,t} = R(\mathbf{q}, \mathbf{o}_i) - \text{mean}(\{R(\mathbf{q}, \mathbf{o}_1), \dots, R(\mathbf{q}, \mathbf{o}_G)\}).$$

Trend 3 — What actually improves RLVR?

Despite lower output length, benchmark scores are the same → Output Length does not really matter.



Trend 3 — What actually improves RLVR?

What about the Reward Functions? Does changing them affect performance?

Spurious Rewards: Rethinking Training Signals in RLVR

**Rulin Shao^{1*} Shuyue Stella Li^{1*} Rui Xin^{1*} Scott Geng^{1*} Yiping Wang¹
Sewoong Oh¹ Simon Shaolei Du¹ Nathan Lambert² Sewon Min³ Ranjay Krishna^{1,2}
Yulia Tsvetkov¹ Hannaneh Hajishirzi^{1,2} Pang Wei Koh^{1,2} Luke Zettlemoyer¹**

¹University of Washington ²Allen Institute for Artificial Intelligence

³University of California, Berkeley

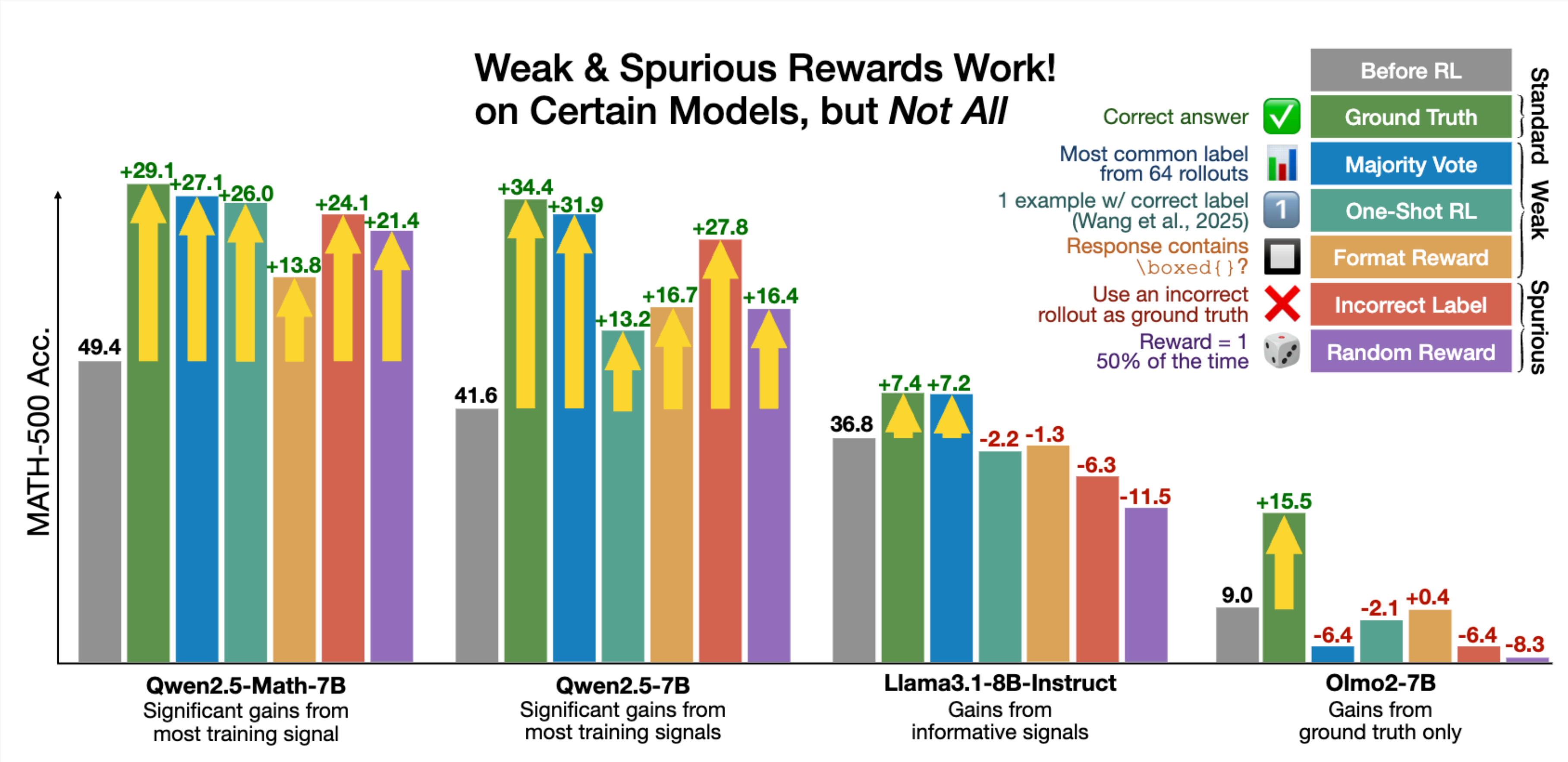
{rulins,stellli,rx31,sgeng}@cs.washington.edu

 [GitHub Repo](#)

Trend 3 – What actually improves RLVR?

What about the Reward Functions? Does changing them affect performance?

Yes, but there are **Failure Modes** (Qwen).



Trend 3 — What actually improves RLVR?

What about the **Objective Functions**?

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) - \beta D_{\text{KL}} \left(\pi_{\theta} \| \pi_{\text{ref}} \right) \right]$$

$$\widehat{A}_i = \frac{R_i - \text{mean} \left(\left\{ R_j \right\}_{j=1}^G \right)}{\text{std} \left(\left\{ R_j \right\}_{j=1}^G \right)}$$

$$r_{i,t}(\theta) = \frac{\pi_{\theta} \left(y_{i,t} \mid x, y_{i,<t} \right)}{\pi_{\theta_{\text{old}}} \left(y_{i,t} \mid x, y_{i,<t} \right)}$$

Trend 3 — What actually improves RLVR?

What about the **Objective Functions**?

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) - \beta D_{\text{KL}} \left(\pi_{\theta} \| \pi_{\text{ref}} \right) \right]$$

$$J_{\text{DAPO}}(\theta) = \frac{1}{\sum_{i=1}^G |y_i|} \sum_{i=1}^G \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon' \right) \widehat{A}_i \right)$$

Trend 3 — What actually improves RLVR?

What about the **Objective Functions**?

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) - \beta D_{\text{KL}} \left(\pi_{\theta} \| \pi_{\text{ref}} \right) \right]$$

$$J_{\text{DAPO}}(\theta) = \frac{1}{\sum_{i=1}^G |y_i|} \sum_{i=1}^G \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon' \right) \widehat{A}_i \right)$$

DAPO — Allows more dynamic sampling, linked to Exploration.

Trend 3 — What actually improves RLVR?

What about the **Objective Functions**? $J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) - \beta D_{\text{KL}} \left(\pi_{\theta} \parallel \pi_{\text{ref}} \right) \right]$

$$J_{\text{GSPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \min \left(s_i(\theta) \widehat{A}_i, \text{clip} \left(s_i(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) \right]$$

$$s_i(\theta) = \left(\frac{\pi_{\theta} (y_i \mid x)}{\pi_{\theta_{\text{old}}} (y_i \mid x)} \right)^{\frac{1}{|y_i|}}$$

DAPO — Allows more dynamic sampling, linked to Exploration.

GSPO — Move clipping to whole sequences to reduce variance and stabilize long generations.

Trend 3 — What actually improves RLVR?

What about the **Objective Functions?**

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x, \{y_i\}} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min \left(r_{i,t}(\theta) \widehat{A}_i, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \widehat{A}_i \right) - \beta D_{\text{KL}} \left(\pi_{\theta} \| \pi_{\text{ref}} \right) \right]$$

$$J_{\text{CISPO}}(\theta) = \mathbb{E} \left[\frac{1}{\sum_{i=1}^G |y_i|} \sum_{i=1}^G \sum_{t=1}^{|y_i|} \text{sg} \left(r_{i,t}(\theta) \right) \widehat{A}_{i,t} \log \pi_{\theta} \left(y_{i,t} \mid x, y_{i,<t} \right) M_{i,t} \right]$$

DAPO — Allows more dynamic sampling, linked to Exploration.

GSPO — Move clipping to whole sequences to reduce variance and stabilize long generations.

CISPO — Clip the importance weight, not the objective for simpler gradients + stable scaling.

‘DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning’ arXiv 2025, DeepSeek-AI

‘DAPO: An Open-Source LLM Reinforcement Learning System at Scale’ NeurIPS 2025, ByteDance

‘Group Sequence Policy Optimization’ arXiv 2025, Qwen

‘MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention’ arXiv 2025, MINIMAX

Trend 3 — What actually improves RLVR?

So do Objective Functions change anything?

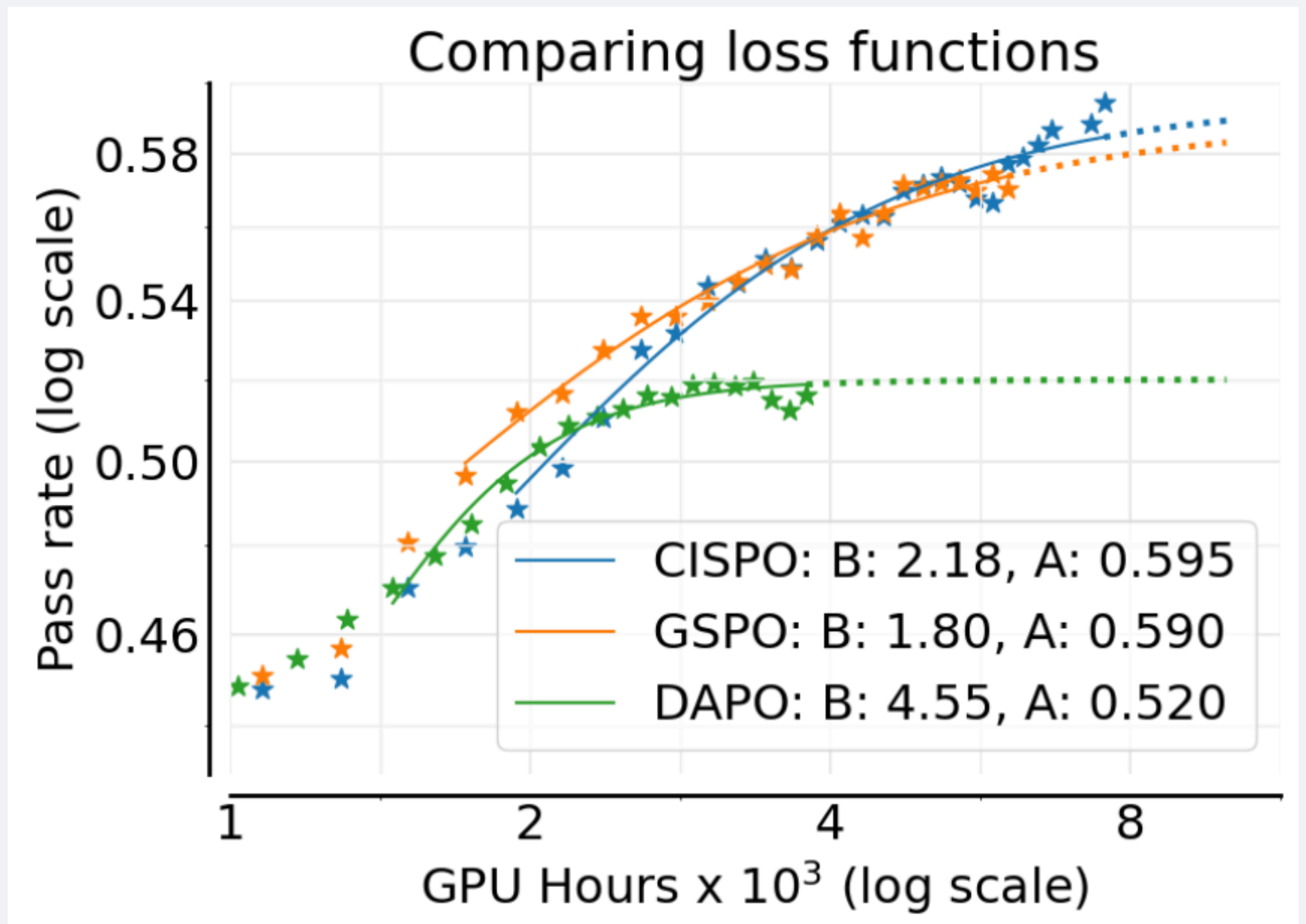
Yes, but we need to view them in the long-term.

The Art of Scaling Reinforcement Learning Compute for LLMs

Devvrit Khatri^{2,*,†}, Lovish Madaan^{1,3,*},
Rishabh Tiwari^{4†}, Rachit Bansal^{5†}, Sai Surya Duvvuri^{2†}, Manzil Zaheer^{1†},
Inderjit S. Dhillon², David Brandfonbrener¹, Rishabh Agarwal^{6,†}

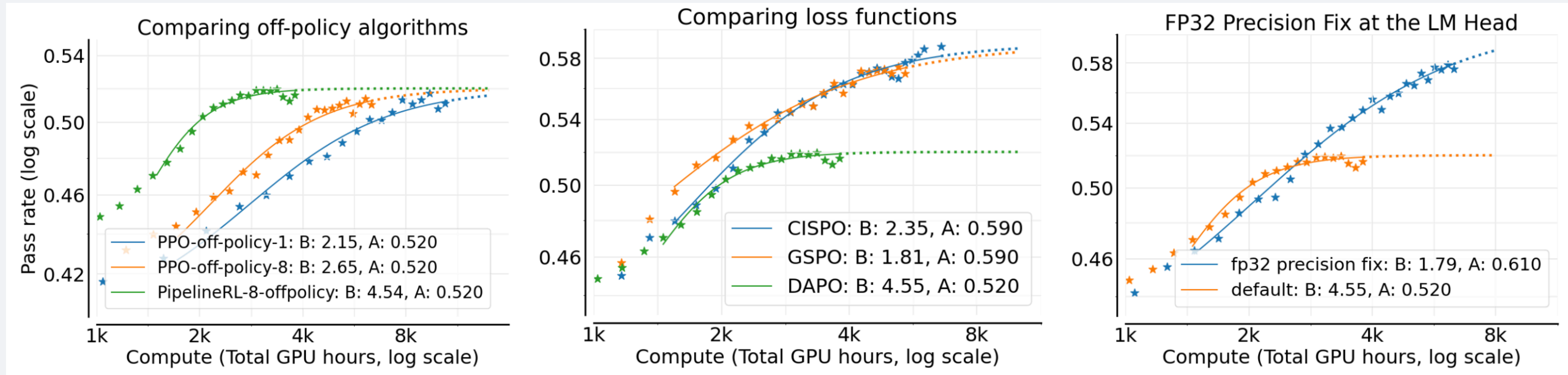
¹Meta, ²UT Austin, ³UCL, ⁴UC Berkeley, ⁵Harvard University, ⁶Periodic Labs

*Equal contribution, [†]Work done at Meta



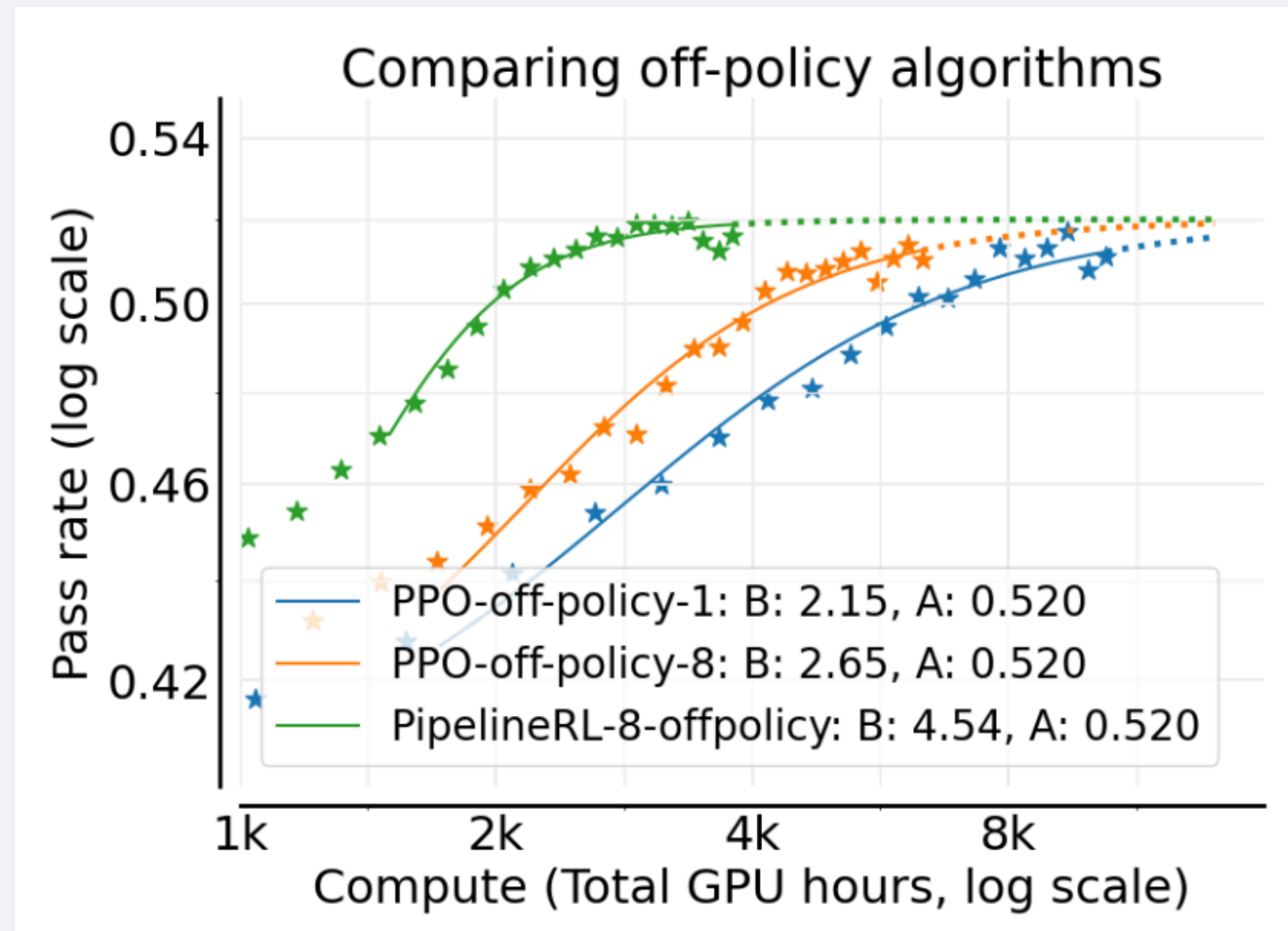
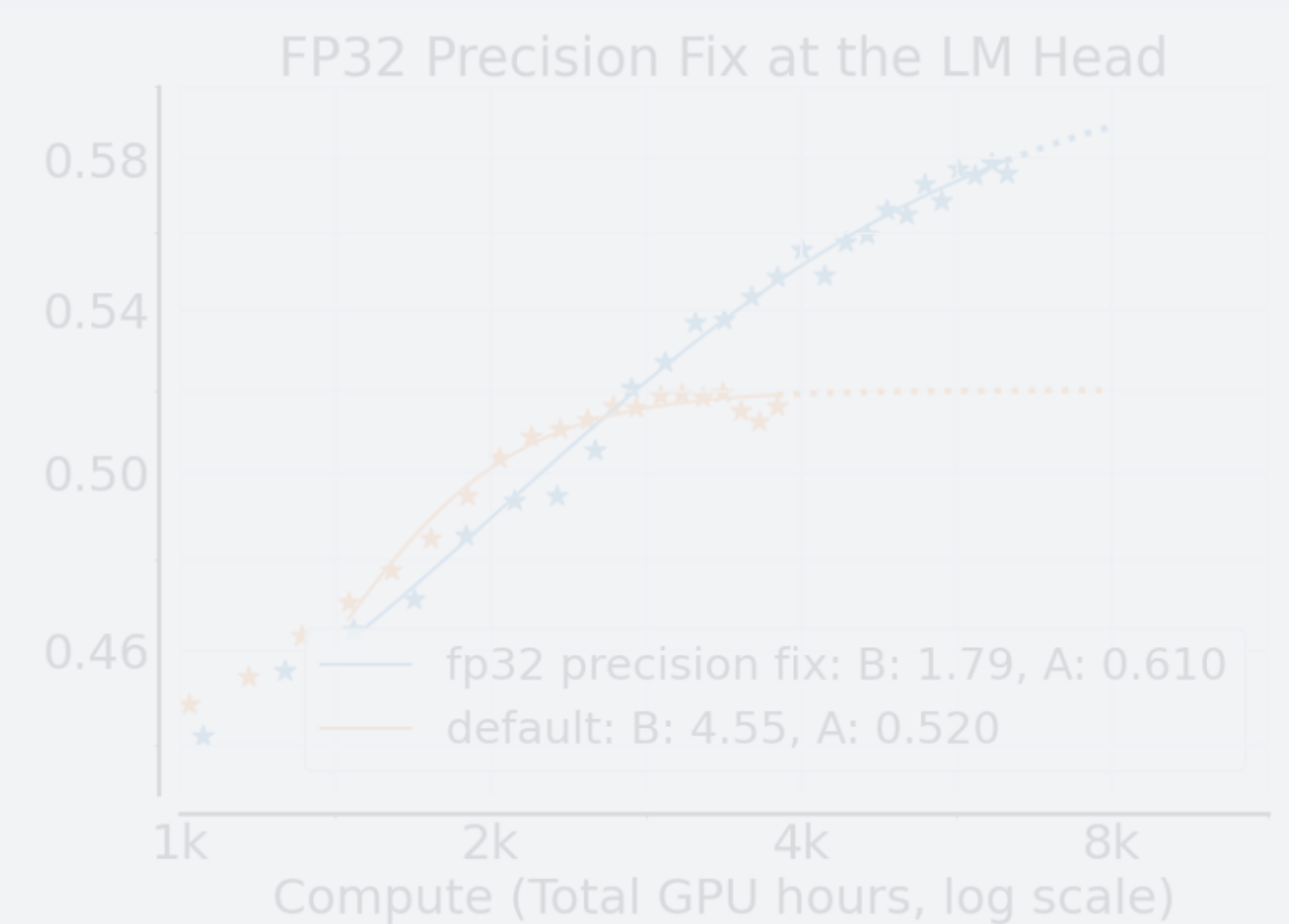
Trend 3 — What actually improves RLVR?

What other factors change in the long-term?



Trend 3 — What actually improves RLVR?

Off-Policy vs On-Policy



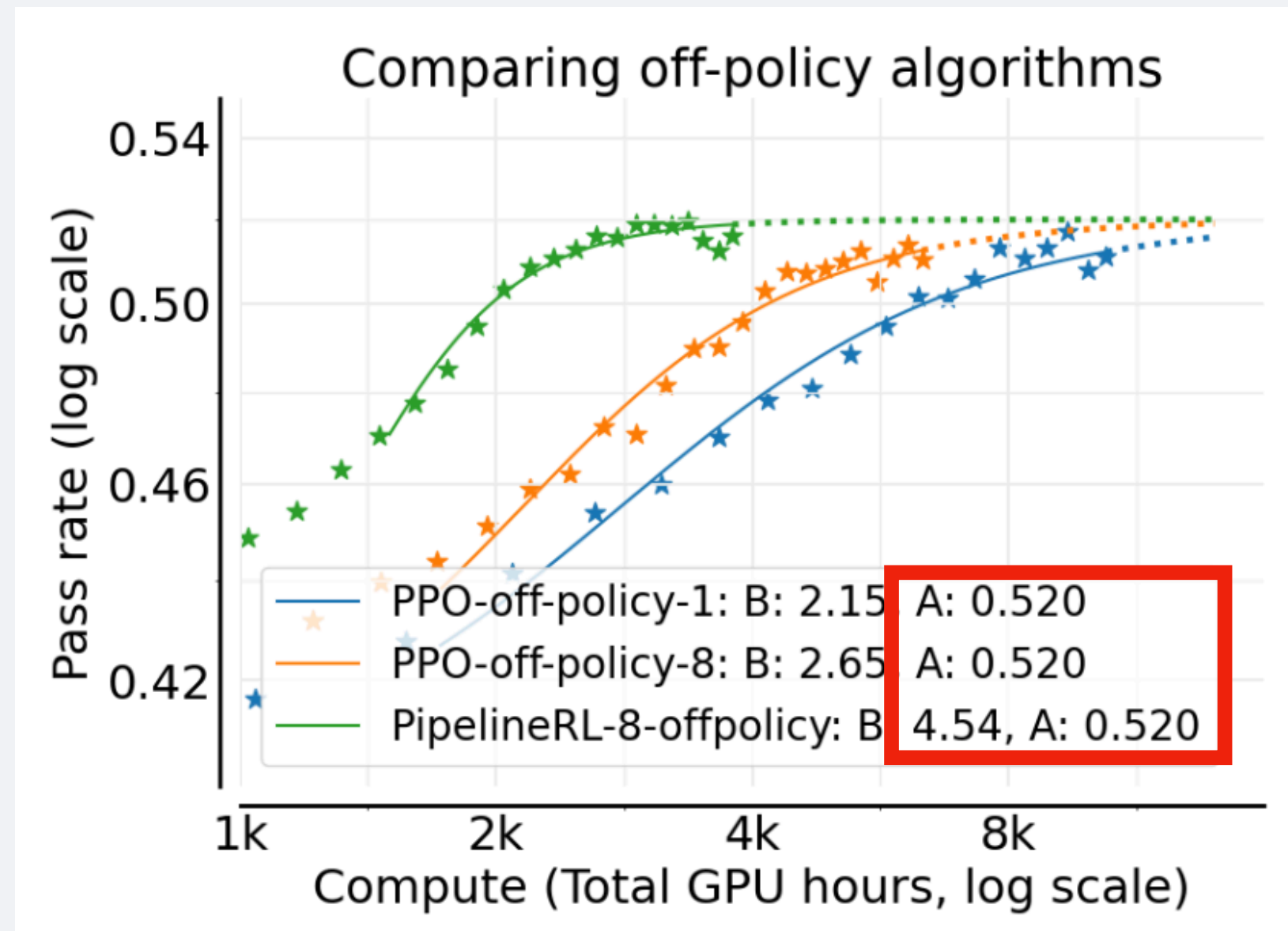
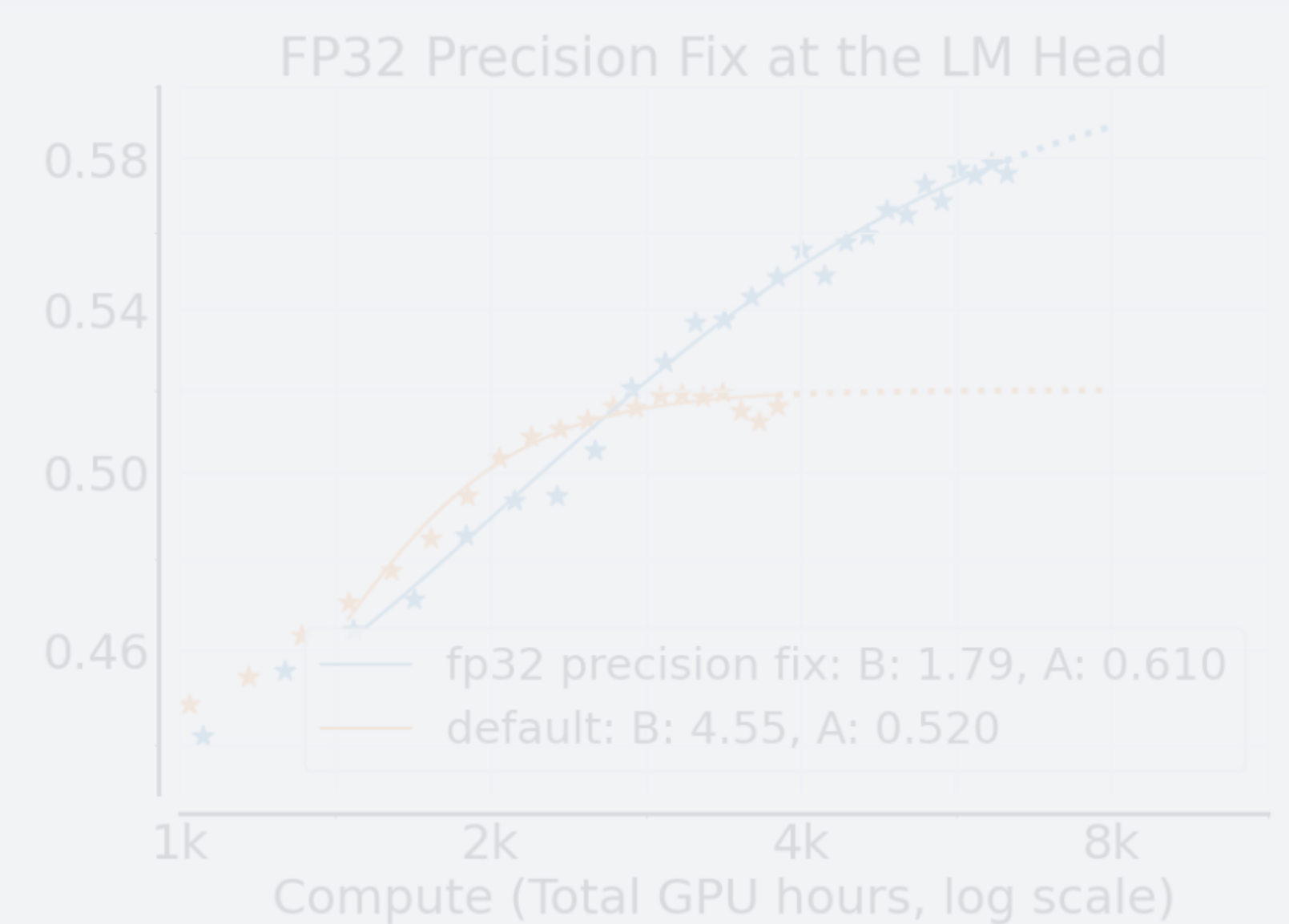
Off-Policy-K — As $K \uparrow$, we do more frequent updates, but the Off-Policy mechanism doesn't change.

PipelineRL — As $K \downarrow$, it is closer to On-Policy.

On-Policy / Off-Policy doesn't really change the limit.

Trend 3 — What actually improves RLVR?

Off-Policy vs On-Policy



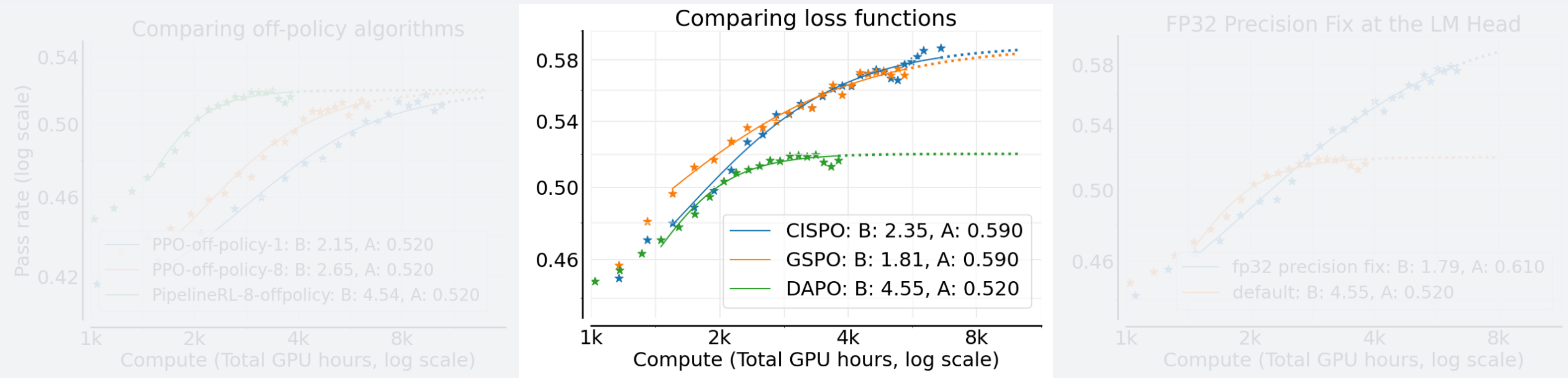
Off-Policy-K — As $K \uparrow$, we do more frequent updates, but the Off-Policy mechanism doesn't change.

PipelineRL — As $K \downarrow$, it is closer to On-Policy.

On-Policy / Off-Policy doesn't really change the limit.

Trend 3 — What actually improves RLVR?

Various Objective Functions



Objective Functions do change the limit, details about them will be ignored in this presentation.

Trend 3 — What actually improves RLVR?

FP32 Precisions



FP32 vs. FP16 matters!

Trend 3 — What actually improves RLVR?

Reward Functions, Objective Functions, FP32 matter in improving RLVR.

→ Improving RLVR is mainly more related to Stable and Safe updates in weights!

Spurious Rewards: Rethinking Training Signals in RLVR

Rulin Shao^{1*} Shuyue Stella Li^{1*} Rui Xin^{1*} Scott Geng^{1*} Yiping Wang¹
Sewoong Oh¹ Simon Shaolei Du¹ Nathan Lambert² Sewon Min³ Ranjay Krishna^{1,2}
Yulia Tsvetkov¹ Hannaneh Hajishirzi^{1,2} Pang Wei Koh^{1,2} Luke Zettlemoyer¹

¹University of Washington ²Allen Institute for Artificial Intelligence

³University of California, Berkeley

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

 [GitHub Repo](#)

The Art of Scaling Reinforcement Learning Compute for LLMs

**Devvrit Khatri^{2,*,†}, Lovish Madaan^{1,3,*},
Rishabh Tiwari^{4†}, Rachit Bansal^{5†}, Sai Surya Duvvuri^{2†}, Manzil Zaheer^{1†},
Inderjit S. Dhillon², David Brandfonbrener¹, Rishabh Agarwal^{6,†}**

¹Meta, ²UT Austin, ³UCL, ⁴UC Berkeley, ⁵Harvard University, ⁶Periodic Labs

*Equal contribution, [†]Work done at Meta

Current Trend

There are a lot of unanswered questions in RLVR, but there are also a lot of findings accomplished in 2025.

People are starting to look deeper into these questions, as well as utilize the findings to make better models / algorithms.

Maximum Likelihood Reinforcement Learning

Fahim Tajwar^{*1} Guanning Zeng^{*2} Yueer Zhou³ Yuda Song¹
Daman Arora¹ Yiding Jiang¹ Jeff Schneider¹ Ruslan Salakhutdinov¹
Haiwen Feng^{4,5} Andrea Zanette¹
¹Carnegie Mellon University ²Tsinghua University ³Zhejiang University
⁴UC Berkeley ⁵Impossible, Inc.
{ftajwar, azanette}@andrew.cmu.edu

Learning to Reason in 13 Parameters

John X. Morris^{1,2}, Niloofar Miresghallah^{1,3}, Mark Ibrahim¹, Saeed Mahlouljifar¹
¹FAIR at Meta, ²Cornell University, ³Carnegie Mellon University

SKILLRL: Evolving Agents via Recursive Skill-Augmented Reinforcement Learning

Peng Xia^{1*} Jianwen Chen^{1*} Hanyang Wang^{1,2*} Jiaqi Liu¹ Kaide Zeng¹ Yu Wang³ Siwei Han¹
Yiyang Zhou¹ Xujiang Zhao⁴ Haifeng Chen⁴ Zeyu Zheng⁵ Cihang Xie⁶ Huaxiu Yao¹

Reinforcement Learning via Self-Distillation

Jonas Hübötter¹ Frederike Lübeck^{*,1,2} Lejs Behric^{*,1} Anton Baumann^{*,1}
Marco Bagatella^{1,2} Daniel Marta¹ Ido Hakimi¹ Idan Shenfeld³
Thomas Kleine Buening¹ Carlos Guestrin⁴ Andreas Krause¹
¹ETH Zurich ²Max Planck Institute for Intelligent Systems ³MIT ⁴Stanford

TEAM OF THOUGHTS: EFFICIENT TEST-TIME SCALING OF AGENTIC SYSTEMS THROUGH ORCHESTRATED TOOL CALLING

Jeffrey T. H. Wong^{1*}, Zixi Zhang^{1*}, Junyi Liu², Yiren Zhao¹
¹Imperial College London, ²Microsoft Research
{tsz.wong20,b.zhang25,a.zhao}@imperial.ac.uk
junyi.liu@microsoft.com

Why Pass@k Optimization Can Degrade Pass@1: Prompt Interference in LLM Post-training

Anas Barakat¹, Souradip Chakraborty², Khushbu Pahwa^{*}, Amrit Singh Bedi³
¹Singapore University of Technology and Design ²University of Maryland, College Park
³University of Central Florida

Summary

2025 Trend :

- Trend 1 — Does RLVR Truly unlock Reasoning?
- Trend 2 — What does RLVR Unlock?
- Trend 3 — What actually improves RLVR?

What can we take away from all of this?

- Although we partially know on what RLVR does, it is still not fully answered if it is worth using it for Reasoning.
- There is not much research yet on actually utilizing the Findings to actually improve RLVR.

감사합니다



통계·데이터사이언스학과 석사 1학기 신재우