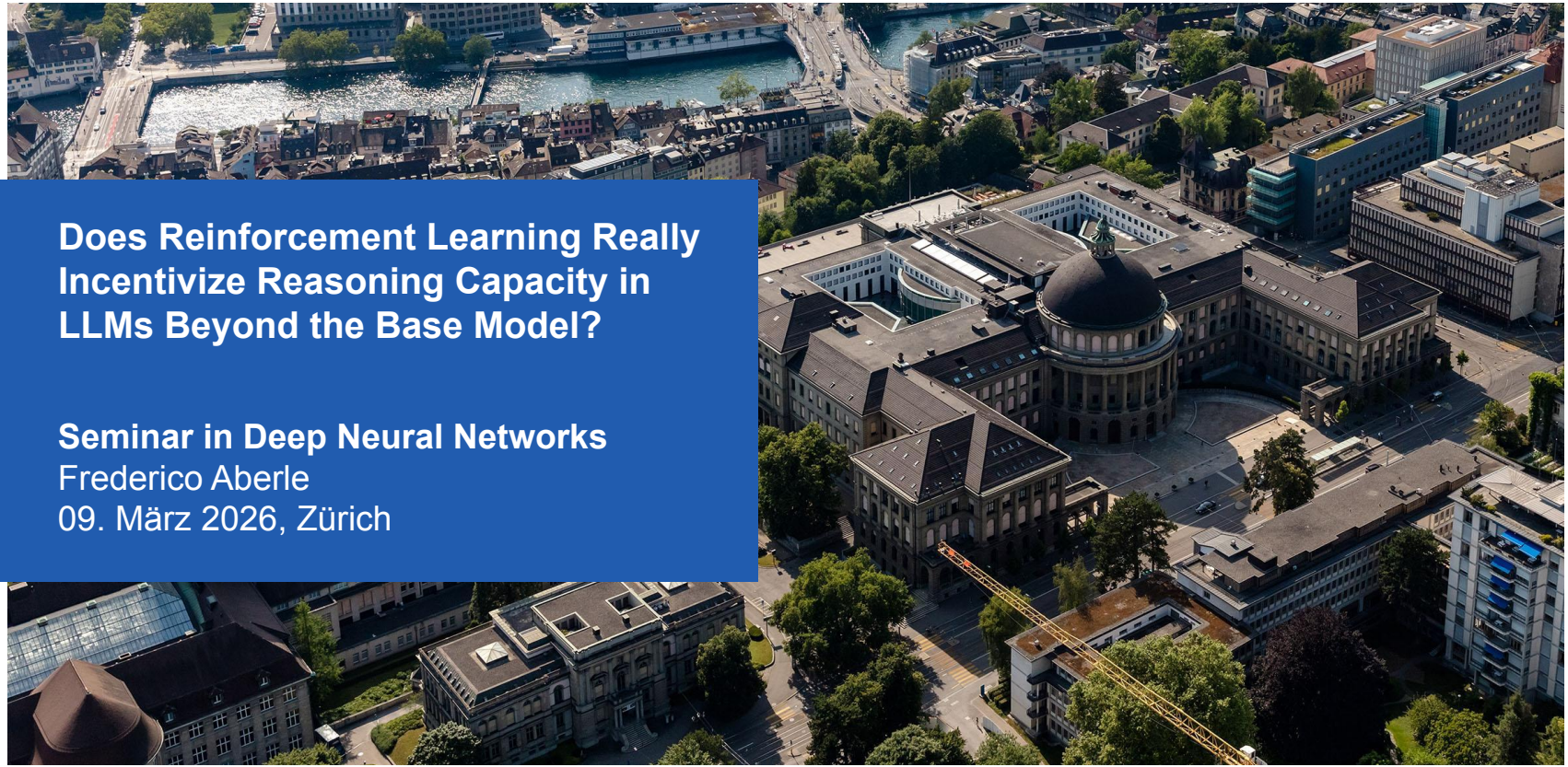
ETH zürich



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

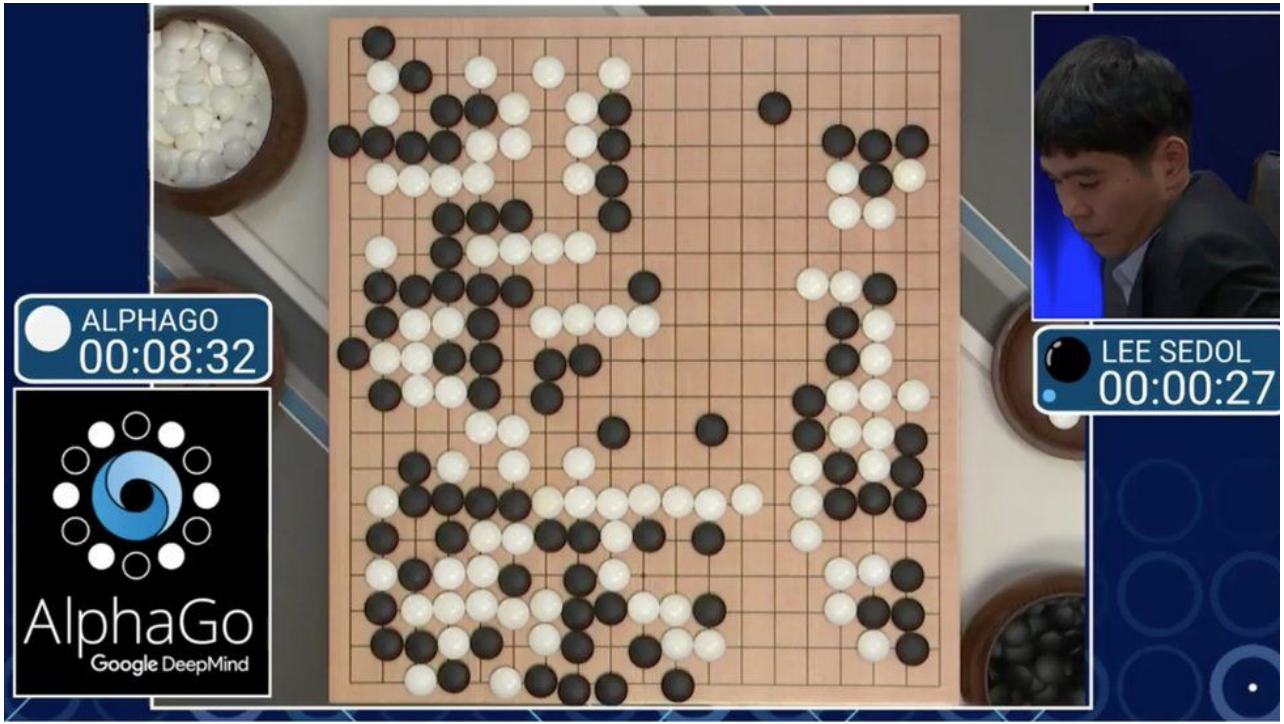
Seminar in Deep Neural Networks

Frederico Aberle

09. März 2026, Zürich

Move 37

https://www.youtube.com/watch?v=WXuK6gekU1Y&t=3018s



https://www.bbc.com/news/technology-35785875

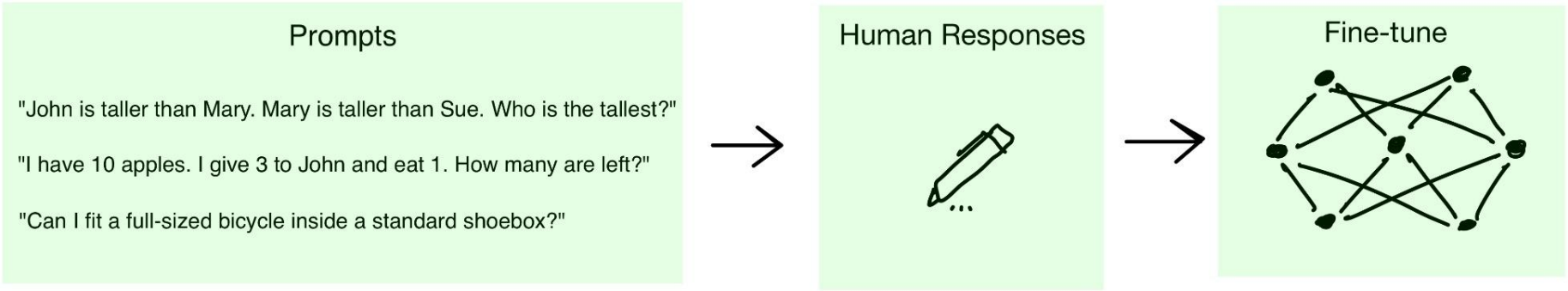
“[...] there was a 1 in 10 000 probability that move 37 would have been played by a human player. [...] It went beyond its human guide [...]”

– David Silver (Lead Researcher on AlphaGo)

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

Frederico Aberle
Mentor: Frédéric Berdoz

Supervised Fine-Tuning (SFT)



elicit reasoning with high-quality responses

Ouyang, Long, Jeff Wu, Xu Jiang, et al., 2022. Training language models to follow instructions with human feedback. arXiv, arXiv:2203.02155.

Supervised Fine-Tuning (SFT)

Logic

Prompts

"John is taller than Mary. Mary is taller than Sue. Who is the tallest?"

Human Responses

Step 1: John > Mary.

Step 2: Mary > Sue.

Step 3: Therefore, John > Mary > Sue.

Answer: John is the tallest.

Supervised Fine-Tuning (SFT)

Prompts

"John is taller than Mary. Mary is taller than Sue. Who is the tallest?"

"I have 10 apples. I give 3 to John and eat 1. How many are left?"

Logic

Arithmetic

Human Responses

Step 1: John > Mary.

Step 2: Mary > Sue.

Step 3: Therefore, John > Mary > Sue.

Answer: John is the tallest.

I started with 10.

Minus 3 for John: 10−3=7.

Minus 1 for me: 7−1=6.

Answer: 6 apples.

Supervised Fine-Tuning (SFT)

Prompts

"John is taller than Mary. Mary is taller than Sue. Who is the tallest?"

"I have 10 apples. I give 3 to John and eat 1. How many are left?"

"Can I fit a full-sized bicycle inside a standard shoebox?"

→

Logic

Arithmetic

Common Sense

Human Responses

Step 1: John > Mary.

Step 2: Mary > Sue.

Step 3: Therefore, John > Mary > Sue.

Answer: John is the tallest.

I started with 10.

Minus 3 for John: 10−3=7.

Minus 1 for me: 7−1=6.

Answer: 6 apples.

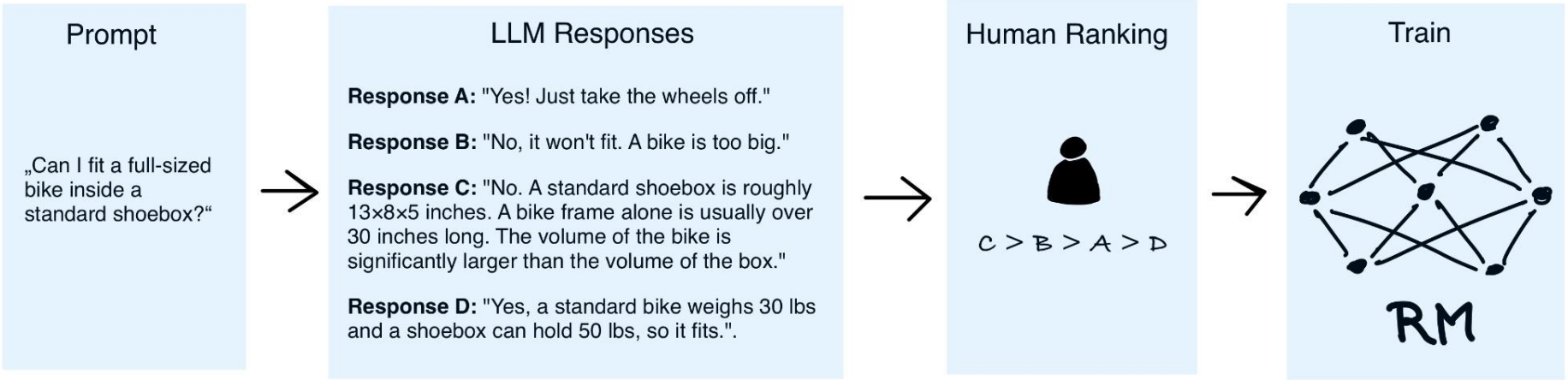
A bicycle is roughly 5 to 6 feet long.

A shoebox is usually about 1 foot long.

Since the bicycle is much larger than the box, it will not fit.

Answer: No.

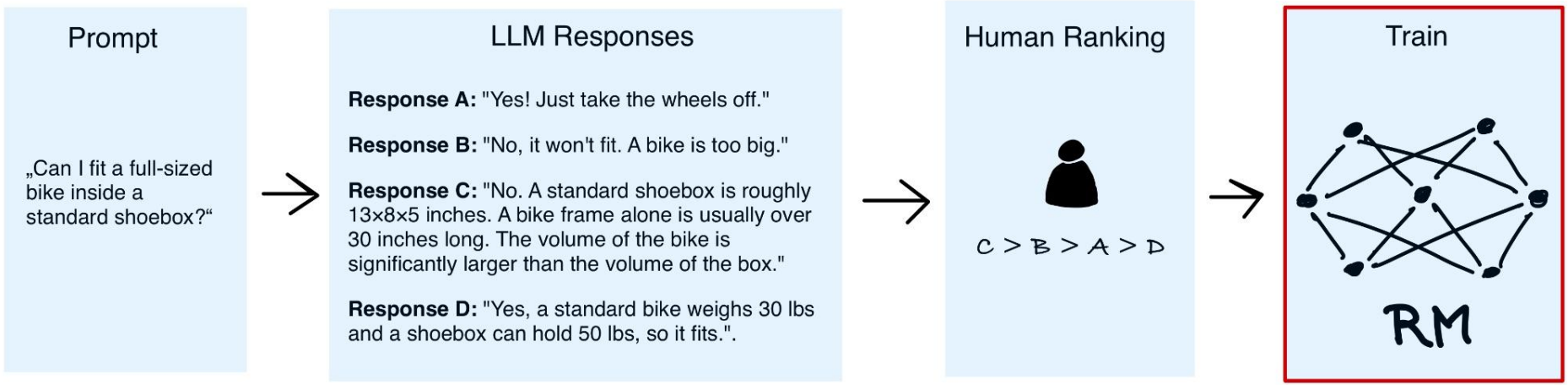
Reinforcement Learning from Human Feedback (RLHF)



scale with reusable Reward Model

Ouyang, Long, Jeff Wu, Xu Jiang, et al., 2022. Training language models to follow instructions with human feedback. arXiv, arXiv:2203.02155.

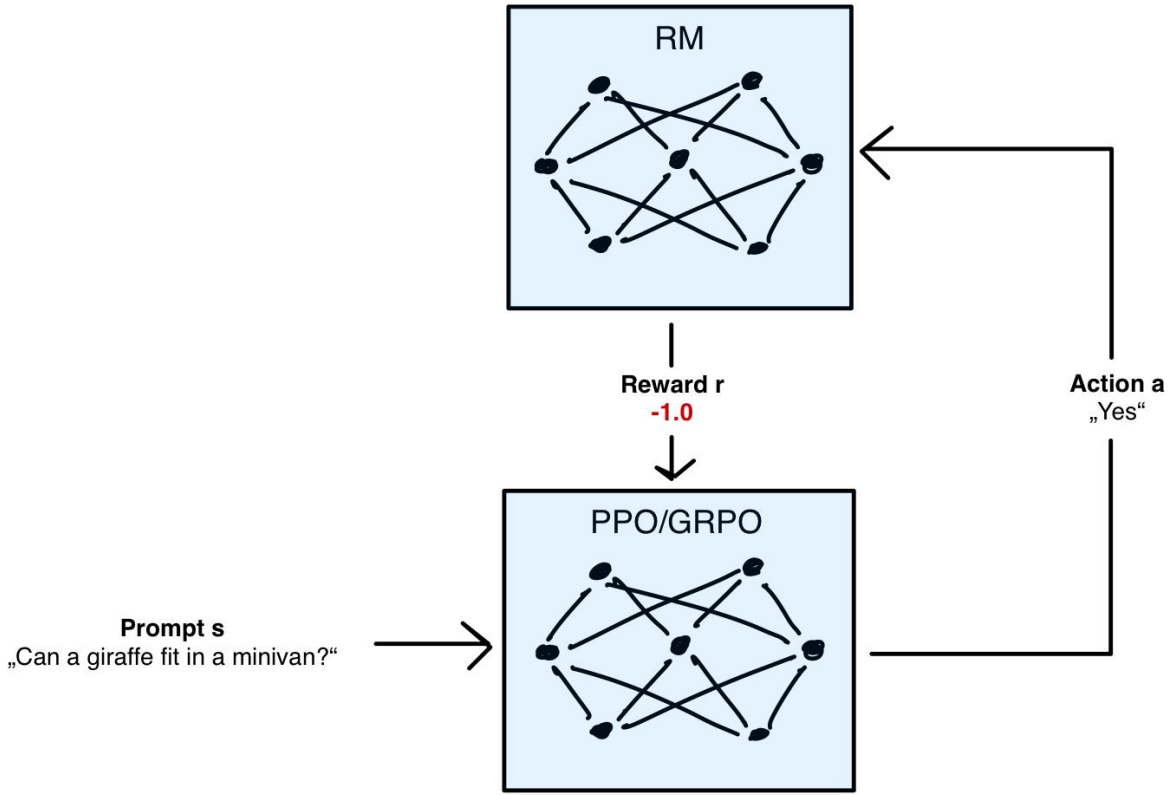
Reinforcement Learning from Human Feedback (RLHF)



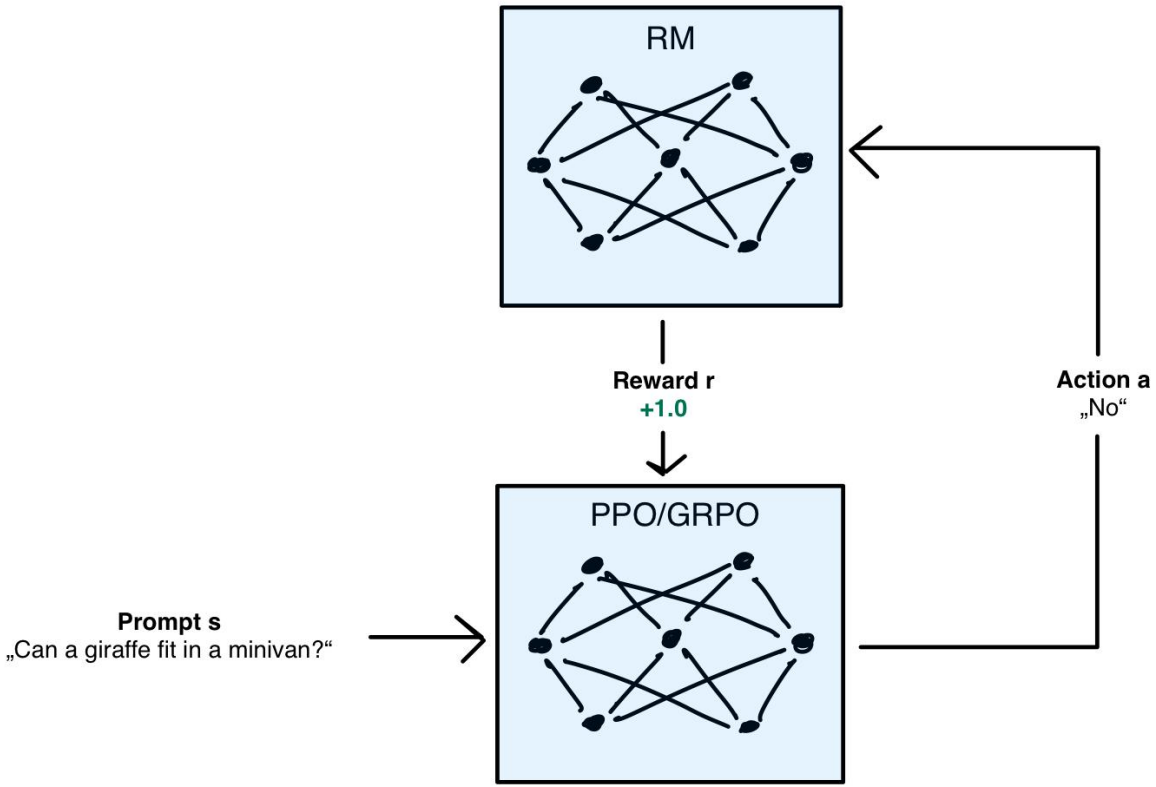
scale with reusable Reward Model

Ouyang, Long, Jeff Wu, Xu Jiang, et al., 2022. Training language models to follow instructions with human feedback. arXiv, arXiv:2203.02155.

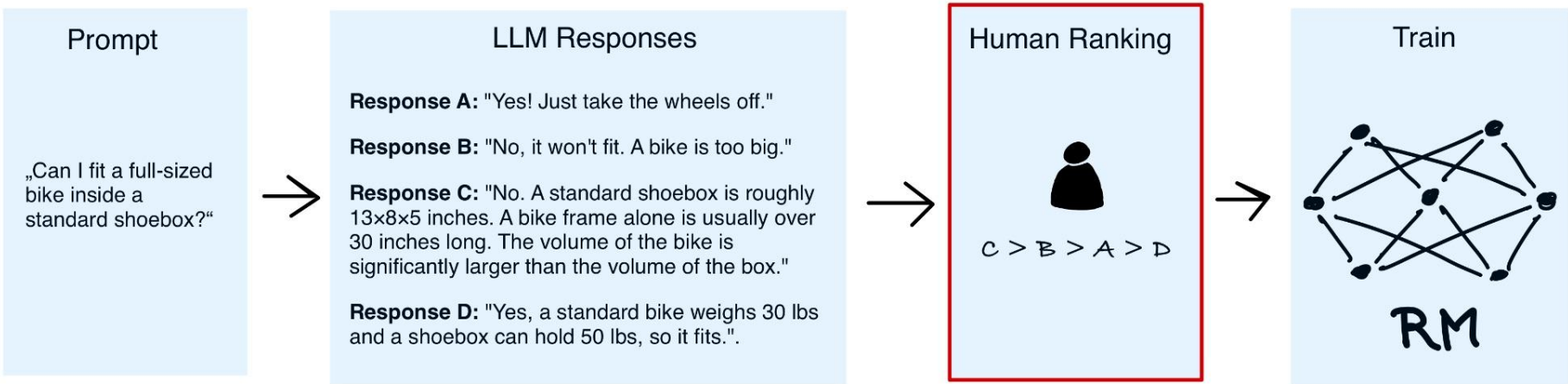
Reinforcement Learning from Human Feedback (RLHF)



Reinforcement Learning from Human Feedback (RLHF)



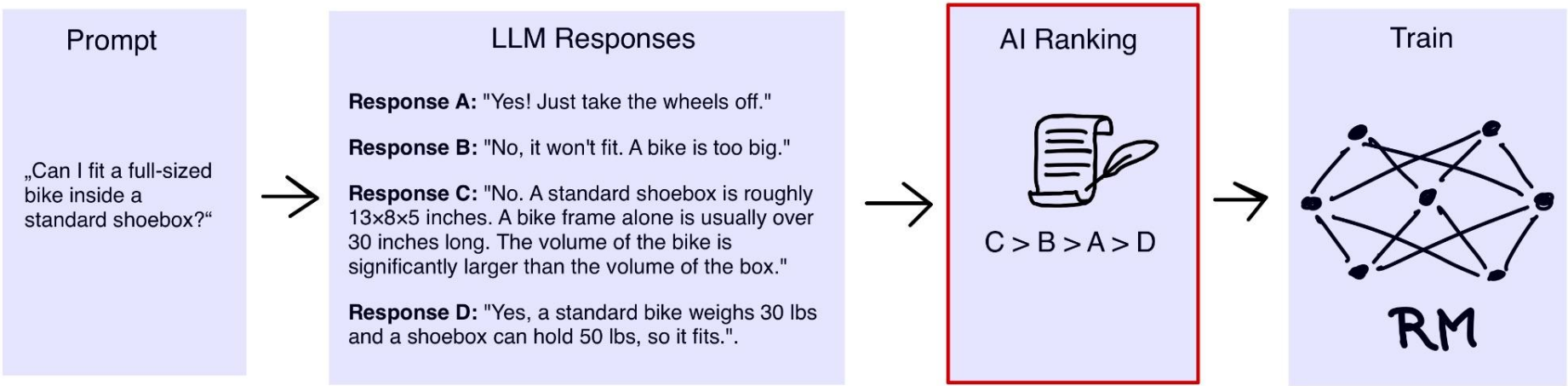
Reinforcement Learning from AI Feedback (RLAIF)



replace bottleneck

Ouyang, Long, Jeff Wu, Xu Jiang, et al., 2022. Training language models to follow instructions with human feedback. arXiv, arXiv:2203.02155.

Reinforcement Learning from AI Feedback (RLAIF)

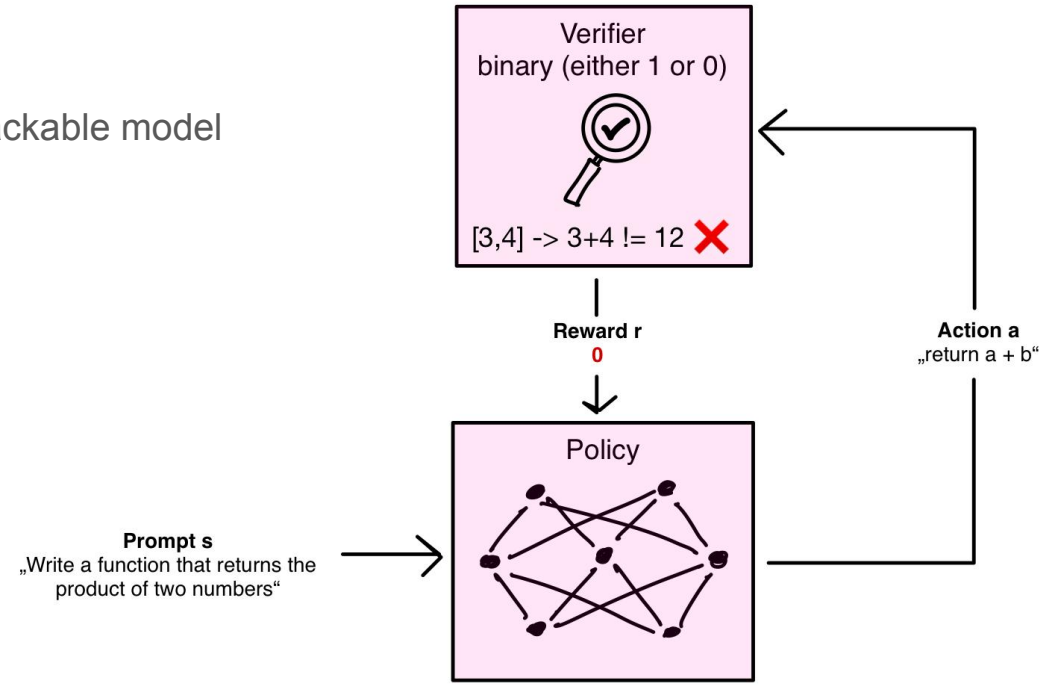


replace bottleneck

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, et al., 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv, arXiv:2212.08073.

Reinforcement Learning with Verifiable Rewards (RLVR)

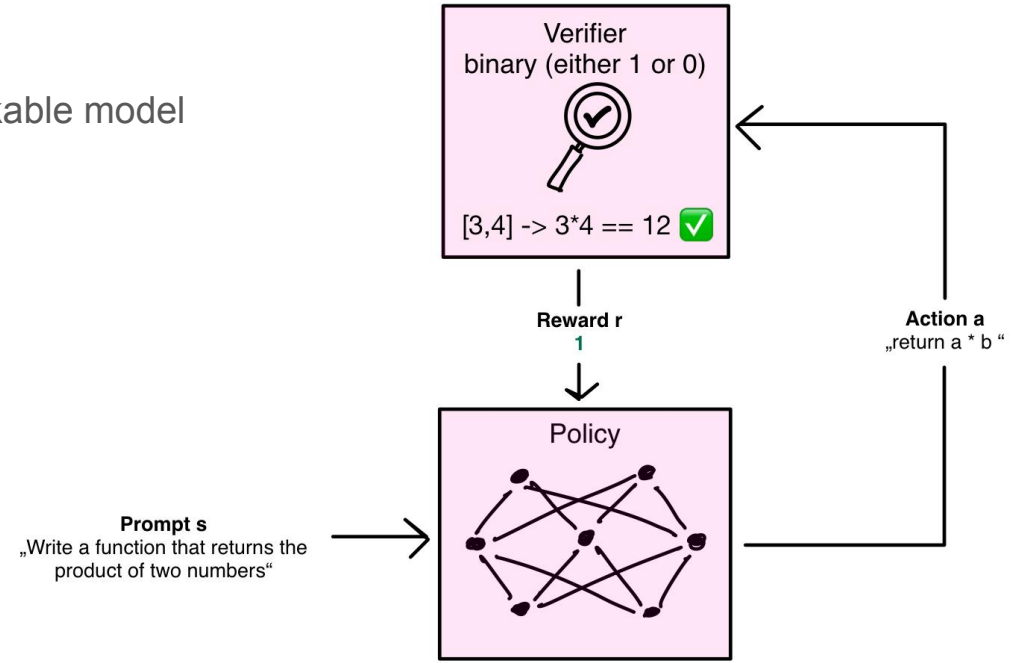
train a non-hackable model



DeepSeek-AI, Guo, Daya, Qihao Zhu, et al., 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv, arXiv:2501.12948.

Reinforcement Learning with Verifiable Rewards (RLVR)

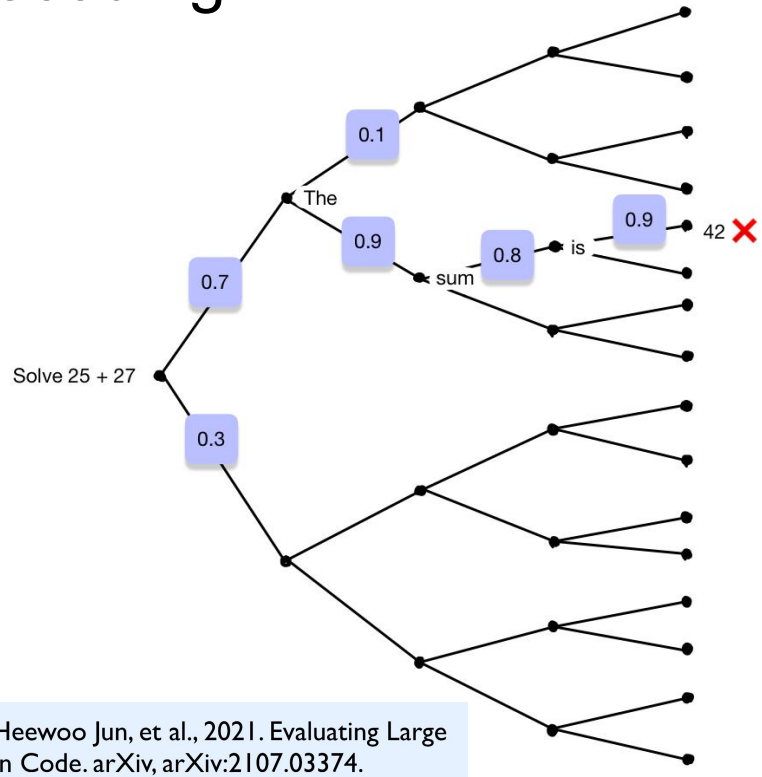
train a non-hackable model



Comparison of Reasoning Methods

Base	SFT	RLHF	RLAIF	RLVR
	HF	HF	AIF	AIF
		Scalable	Scalable	Scalable
		Hackable	Hackable	Aligned

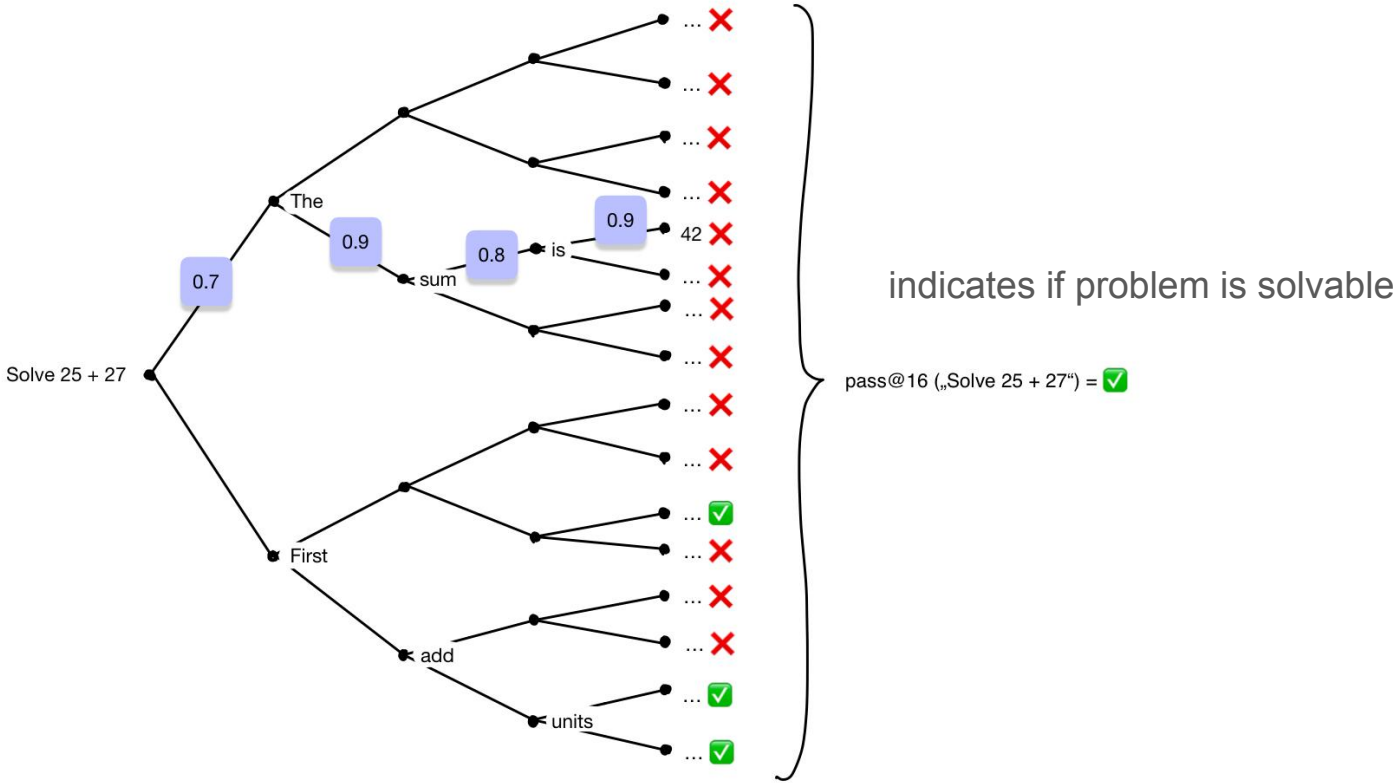
Evaluation metrics for reasoning capabilities - Greedy Decoding



underestimates true potential

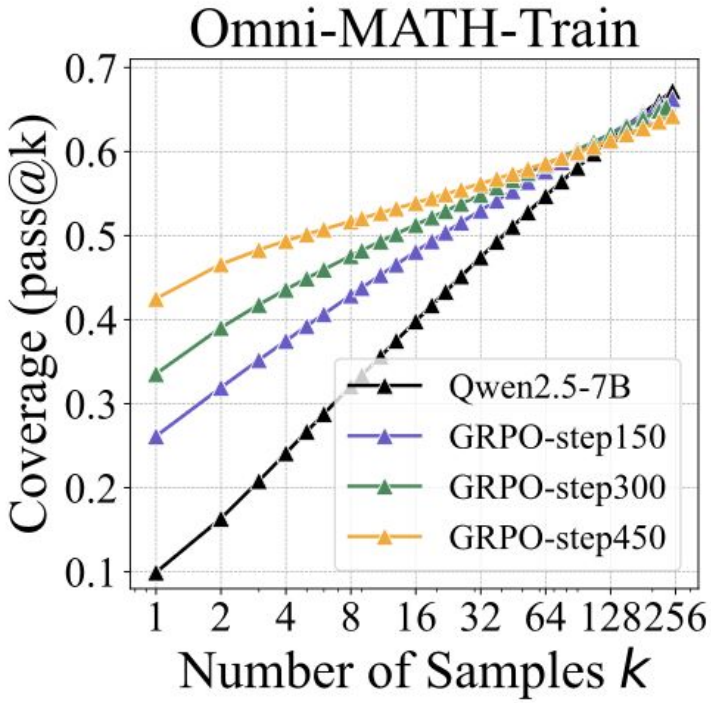
Chen, Mark, Jerry Tworek, Heewoo Jun, et al., 2021. Evaluating Large Language Models Trained on Code. arXiv, arXiv:2107.03374.

Evaluation metrics for reasoning capabilities - pass@k



Does RLVR expand coverage?

narrows reasoning boundary



Yue, Yang, Zhiqi Chen, Rui Lu, et al., 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? arXiv, arXiv:2504.13837.

Does RLVR discover new reasoning strategies?

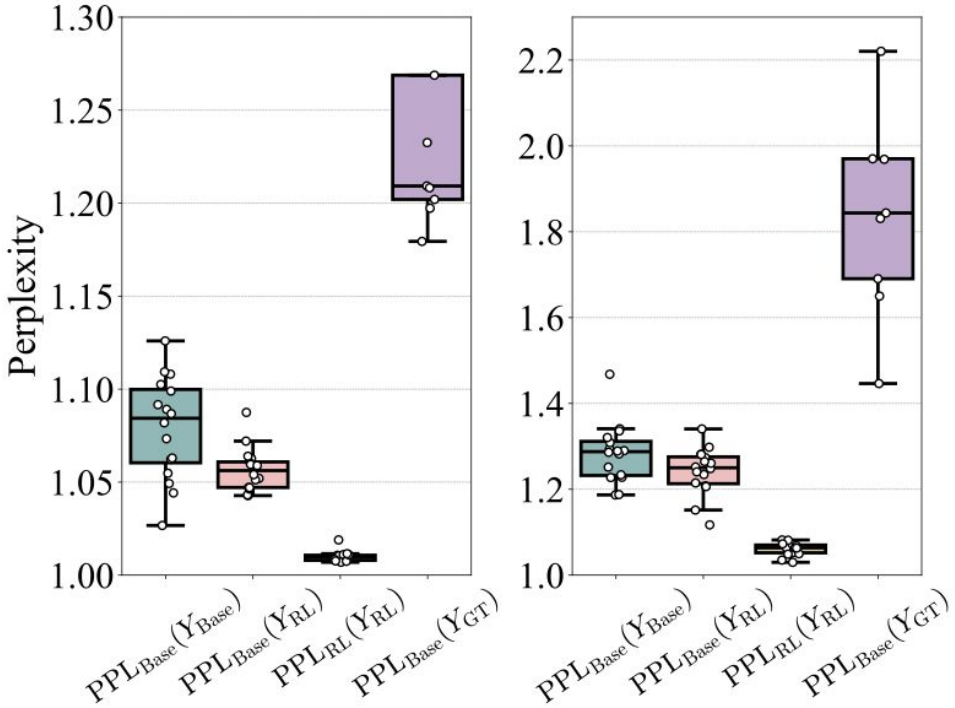
$$\mathrm{PPL}_m(\mathbf{Y} \mid x) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P(y_t \mid x, y_1, \ldots, y_{t-1})\right)$$

Likelihood of model m generating response Y conditioned on prompt x

Yue, Yang, Zhiqi Chen, Rui Lu, et al., 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? arXiv, arXiv:2504.13837.

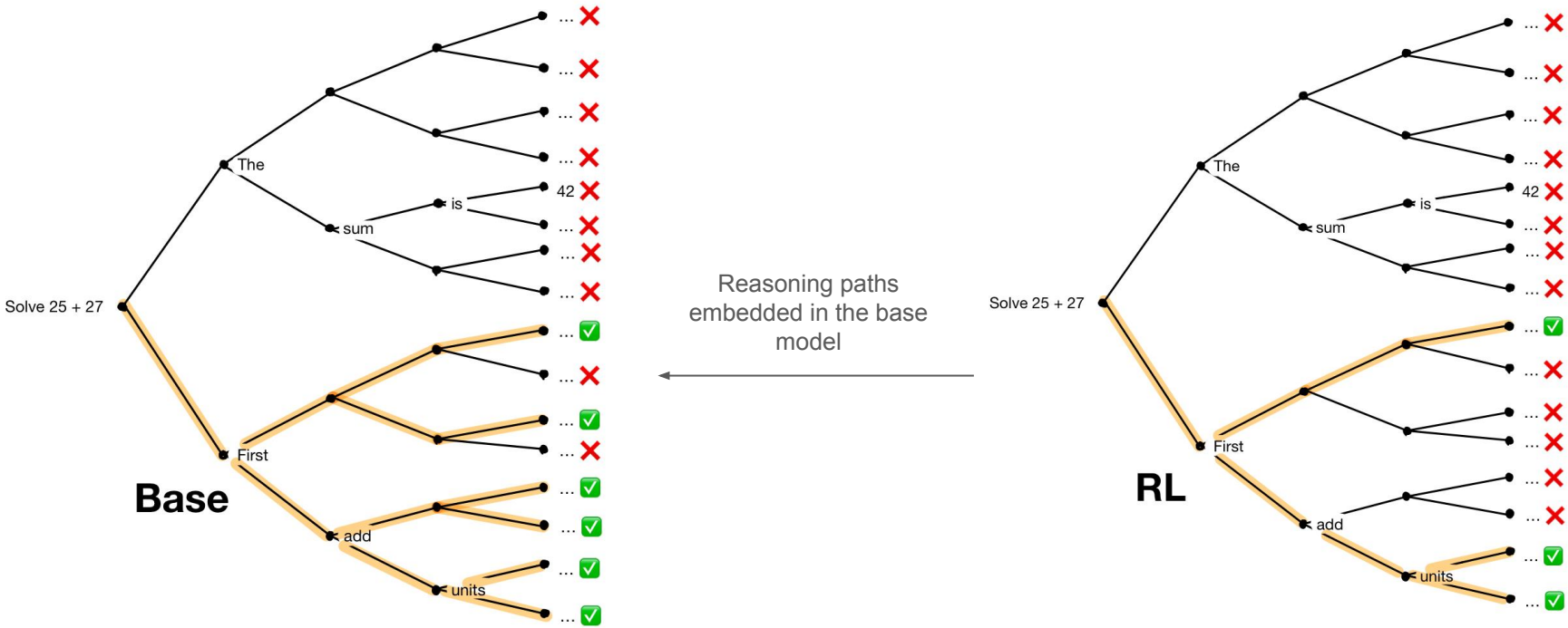
Does RLVR discover new reasoning strategies?

RL dist. embedded in Base dist.



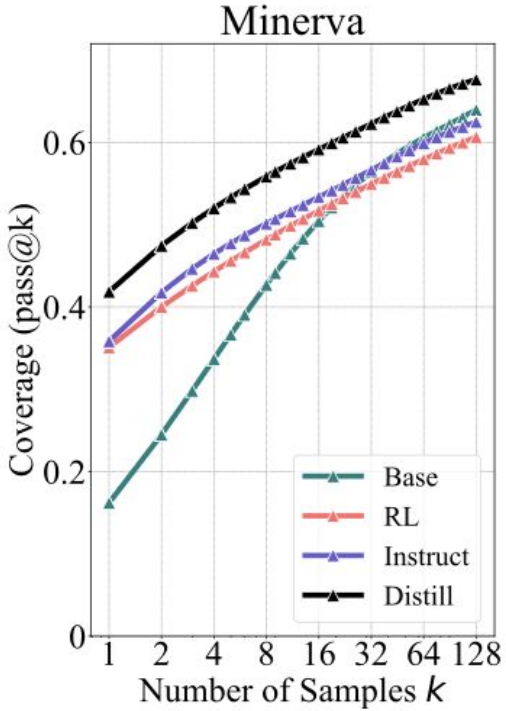
Yue, Yang, Zhiqi Chen, Rui Lu, et al., 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? arXiv, arXiv:2504.13837.

Does RLVR discover new reasoning strategies?



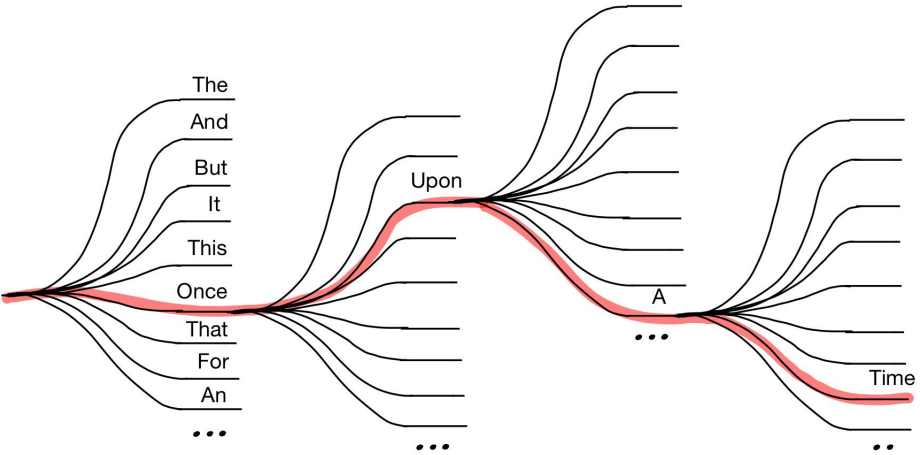
How does RL compare to Distillation?

Distillation can introduce new reasoning patterns



Yue, Yang, Zhiqi Chen, Rui Lu, et al., 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? arXiv, arXiv:2504.13837.

Discussion - Why is RLVR bounded by the base model?



exponentially larger action space

A bicycle is roughly 5 to 6 feet long. +0,1

A shoebox is usually about 1 foot long. + 1,2

Since the bicycle is much larger than the box, it will not fit. +1,5

Answer: No. +5

introduce intermediate instead of binary process rewards

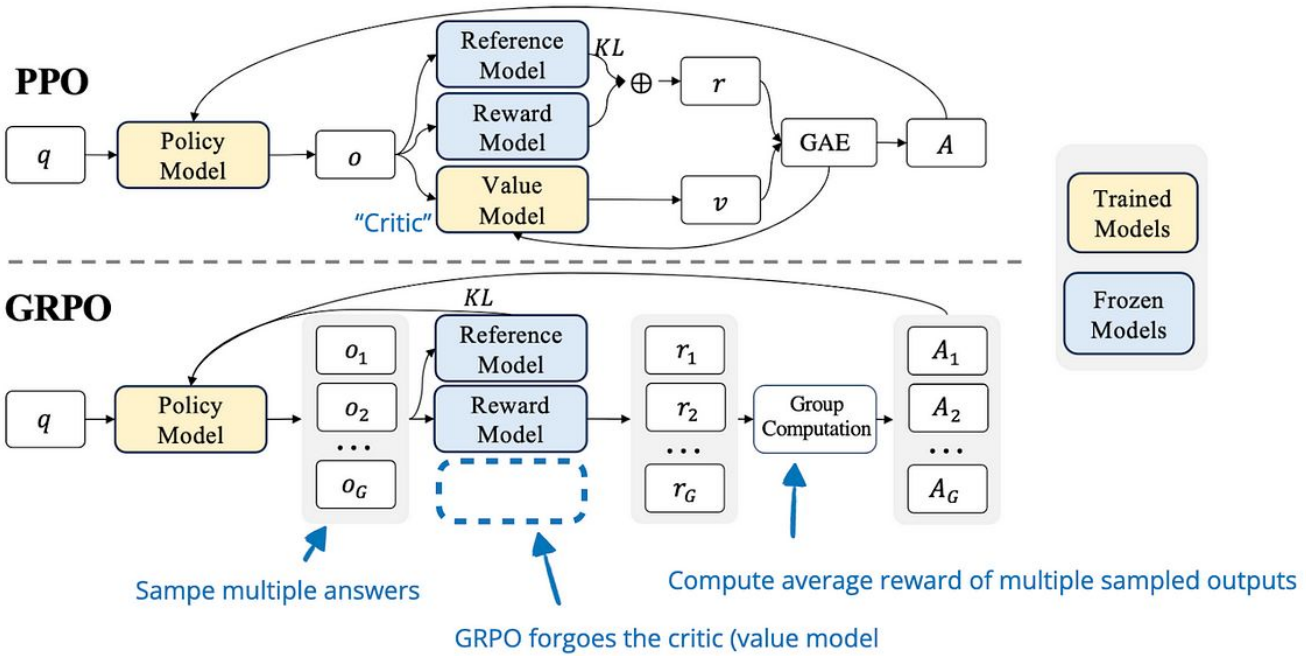
Conclusion

- Reasoning capabilities of RLVR-trained models remain bounded by the base model
- Current RLVR methods have not fully realized the potential to elicit novel reasoning abilities

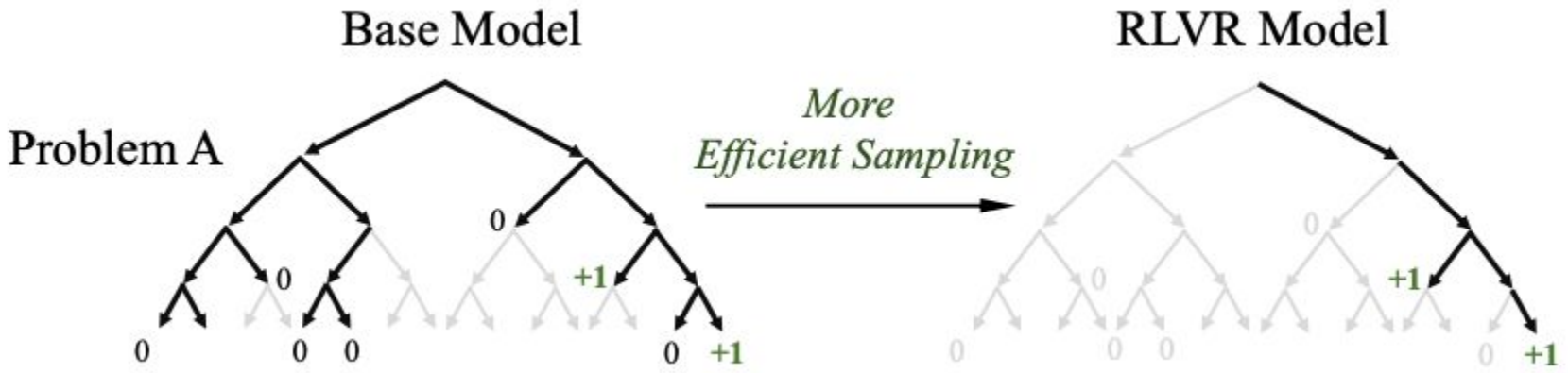
My opinion

- RL is all you need for reasoning is just a hype
- Rigorous work (benchmarking across models and training methods)
- Neglects subjective reasoning (e.g. creative writing)
- Work only conducted on current open-source models

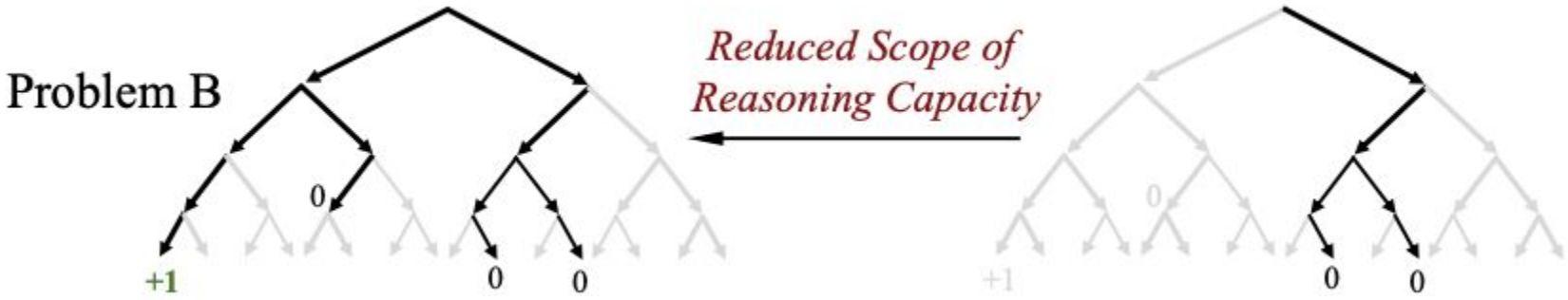
PPO & GRPO


PPO
q
Policy Model
o
Reference Model
Reward Model
Value Model
KL
⊕
r
v
GAE
A
"Critic"
GRPO
KL
q
Policy Model
o_1
o_2
o_G
Reference Model
Reward Model
r_1
r_2
r_G
Group Computation
A_1
A_2
A_G
Trained Models
Frozen Models
Sampe multiple answers
GRPO forgoes the critic (value model
Compute average reward of multiple sampled outputs

1st Finding - Sampling Efficiency



2nd Finding - Narrowing Reasoning Capability Boundary



Evidence with Accuracy Distribution Analysis

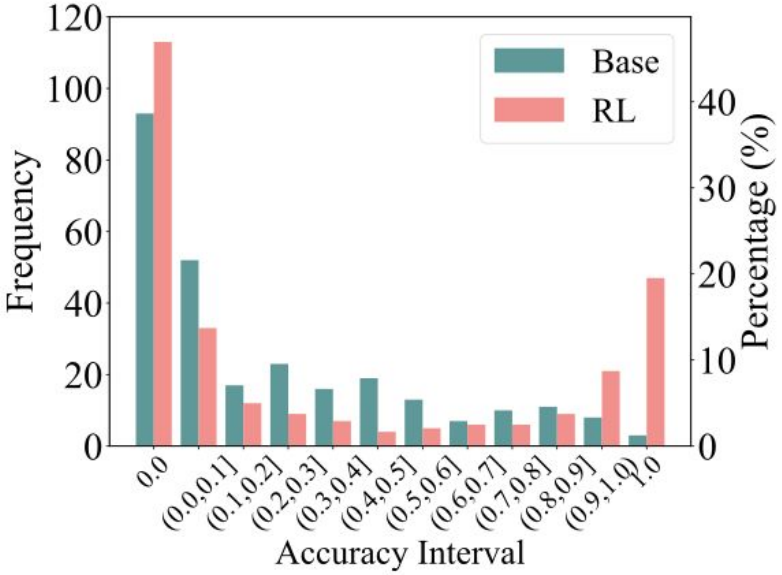


Figure 5: Qwen2.5-7B Accuracy Histogram on Minerva.

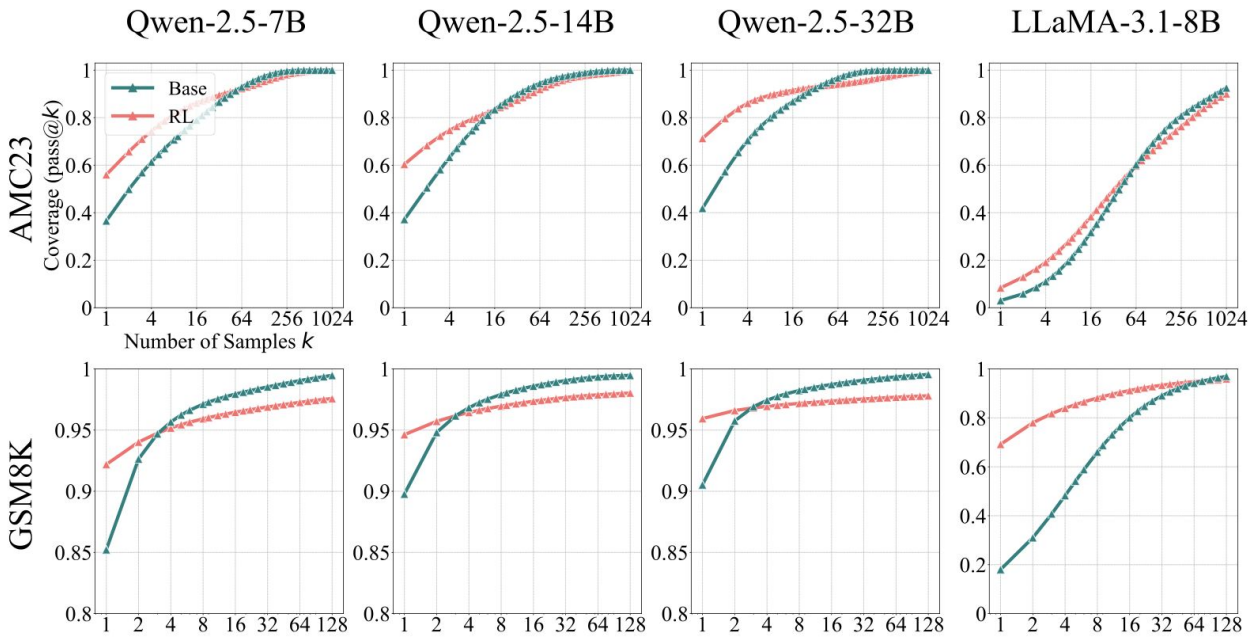


Figure 10: More results of SimpleRLZoo on GSM8K and AMC23.