Reading is fun

"Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., & Huang, G. (2025). Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?"

Chong Cher

July 16, 2025

Outline

SILE Office Access

Thanks to SILE for permission to use the premises for the reading group.

All participants of the reading group are to enter from the main door, and are only allowed to use the designated meeting room.

Please do not access the main office.

Havelock 2 Washroom

The washroom is located outside of the office, on the opposite end of the lift lobby.

Please press the doorbell once you are at the main door; someone in the meeting room will open the door for you.

Chatham House Rule

The reading sessions will be held under Chatham House Rule —

> *When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed.*

Feel free to share ideas openly, but please respect confidentiality.

Reading Group Expectations

This is a community-led reading group broadly interested in Machine Learning / Artificial Intelligence topics. Some simple guidelines:

- Treat others as you would want to be treated
- Offensive language is prohibited
- Read (or skim) the paper; the discussion is much more productive if we have a common baseline for discussion
- Please volunteer to present; this could be a paper you find interesting, or even a project you would like feedback on or to discuss.

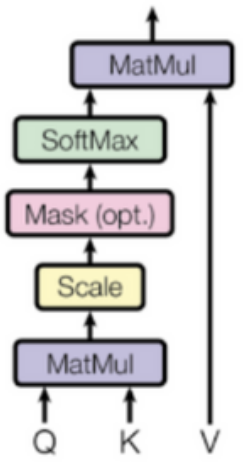
How to Contribute

Please feel free to use the Telegram group chat ("Discussion is fun") to discuss the presented paper, or any interesting AI/ML news you may have come across.

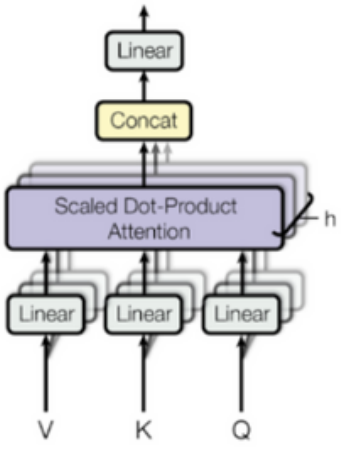
If you would like to present, please contact me (@cheekycheeky on Telegram) with a brief description of the topic.

Once I have approved it, I will announce the next session (typically Thursdays) on the Telegram channel ("Reading is fun"), ideally one month in advance.

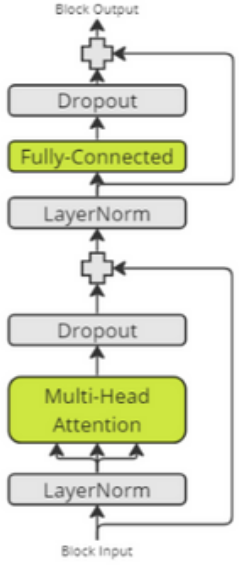
Transformers



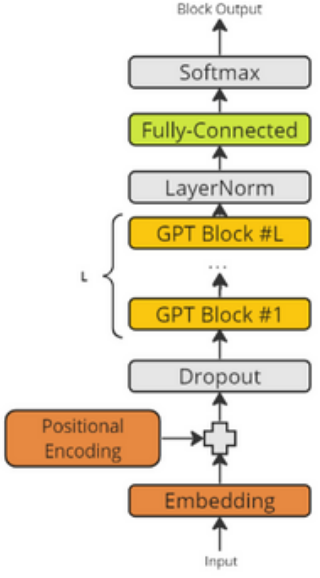
Scale Causal Attention



Multi-Head Attention

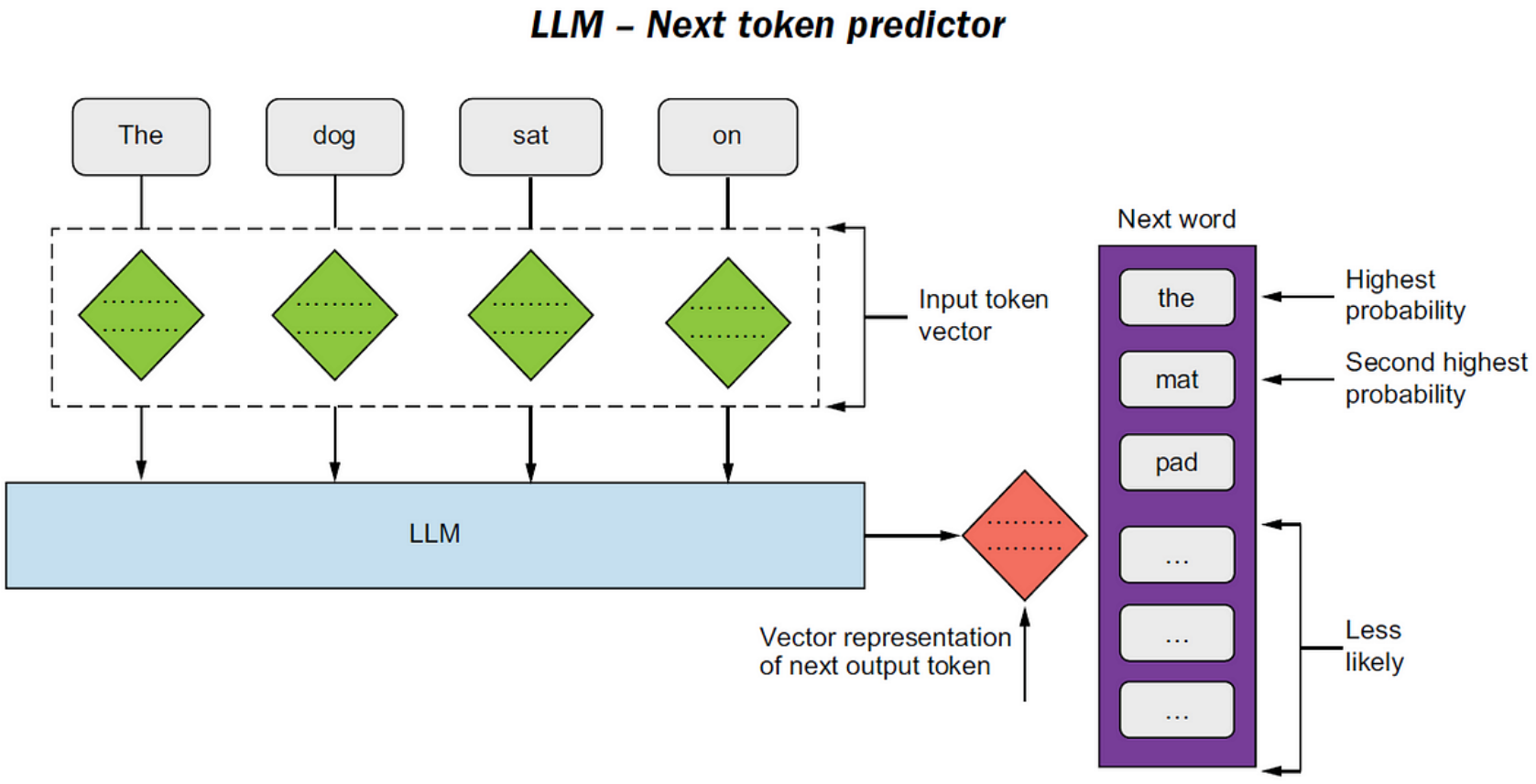


Transformer Block

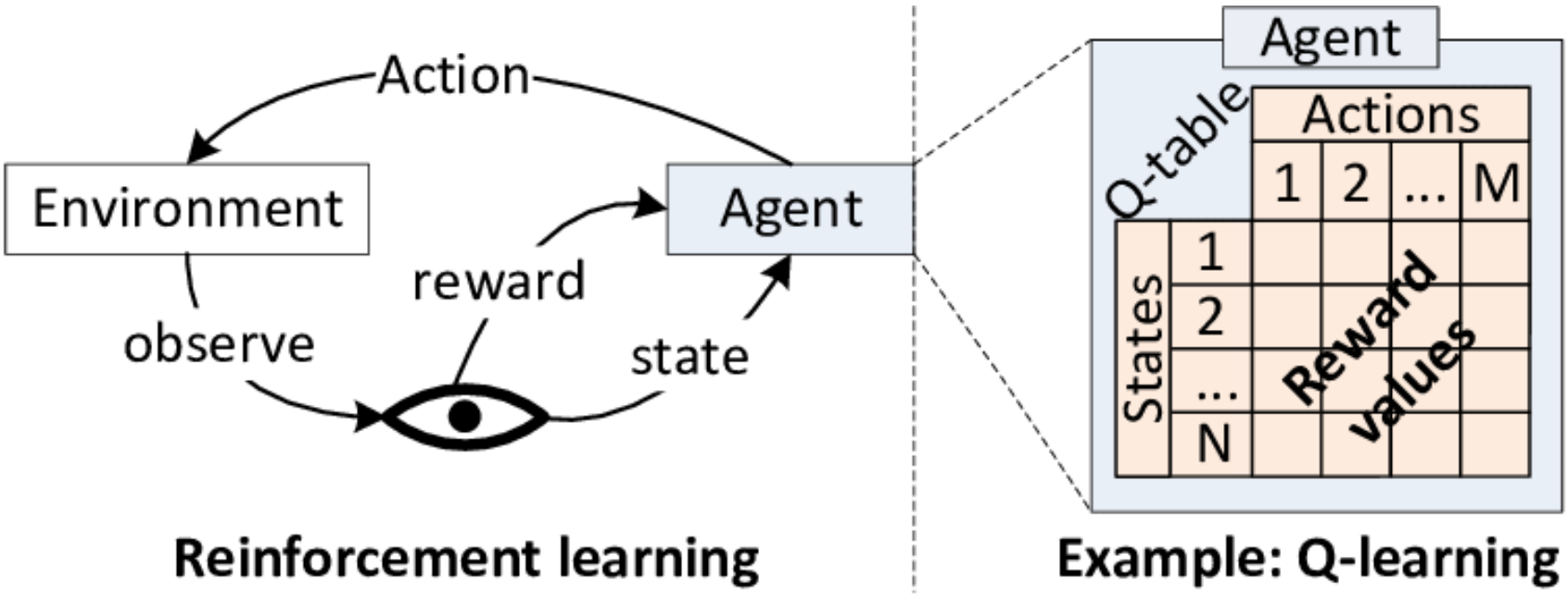


GPT Model

Language Model



Reinforcement Learning



Example: RLHF (GPT-3)

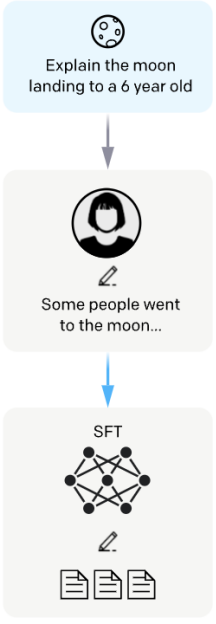
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



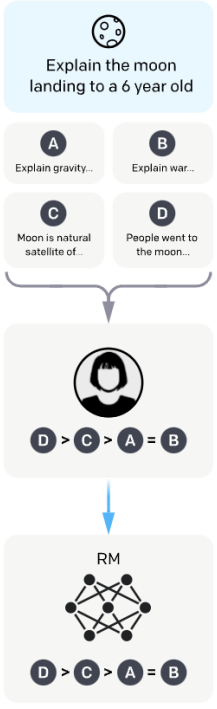
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

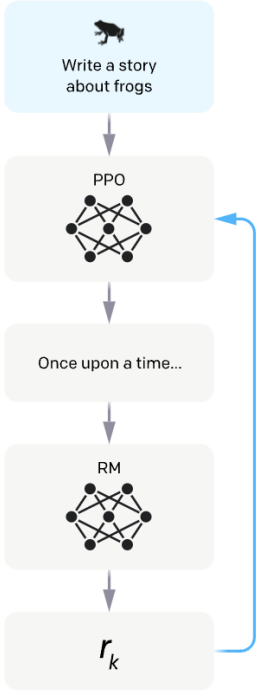
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



See also: Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.

Problems with RLHF?

- ???

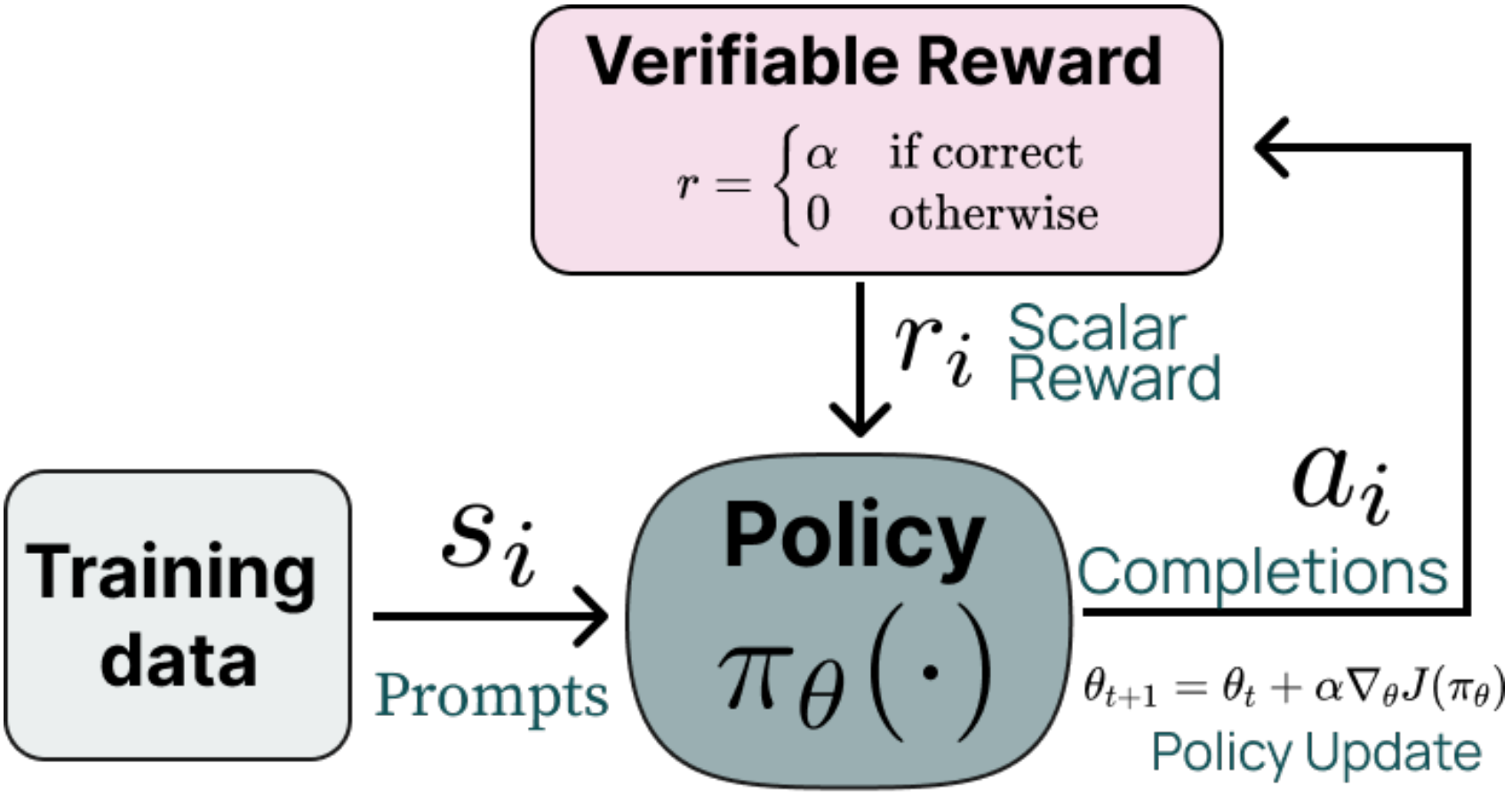
Problems with RLHF?

- ???
- What is the limiting factor for RLHF?

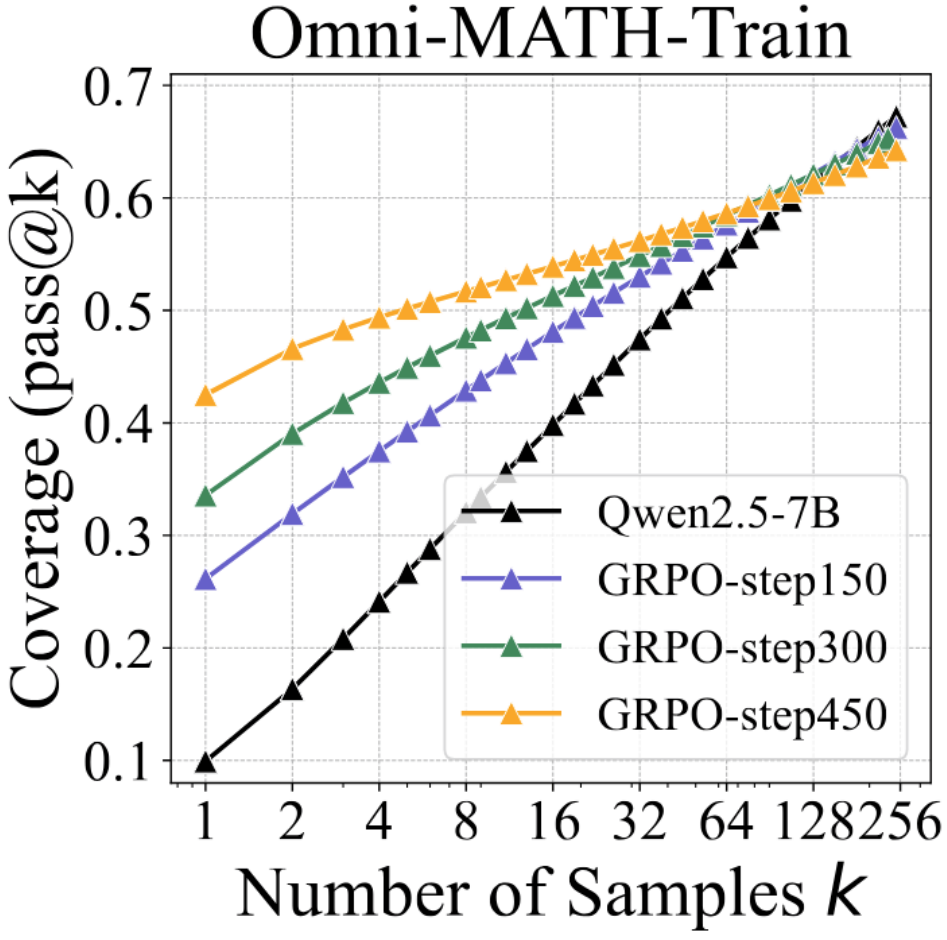
Problems with RLHF?

- ???
- What is the limiting factor for RLHF?
- How can we scale it up?

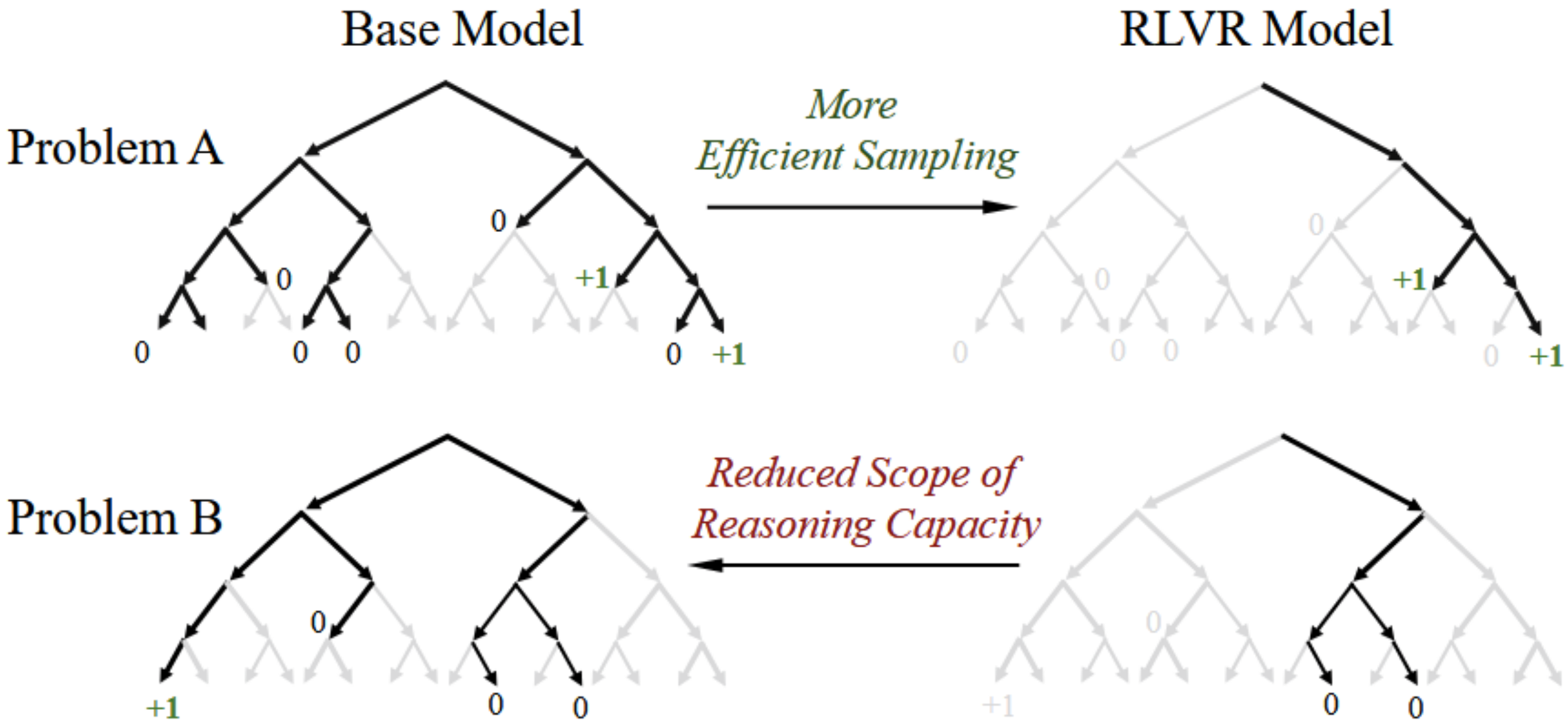
Verifiable Rewards



Effect of current RLVR on LLM's reasoning ability



Paper's hypothesis



Pass@K

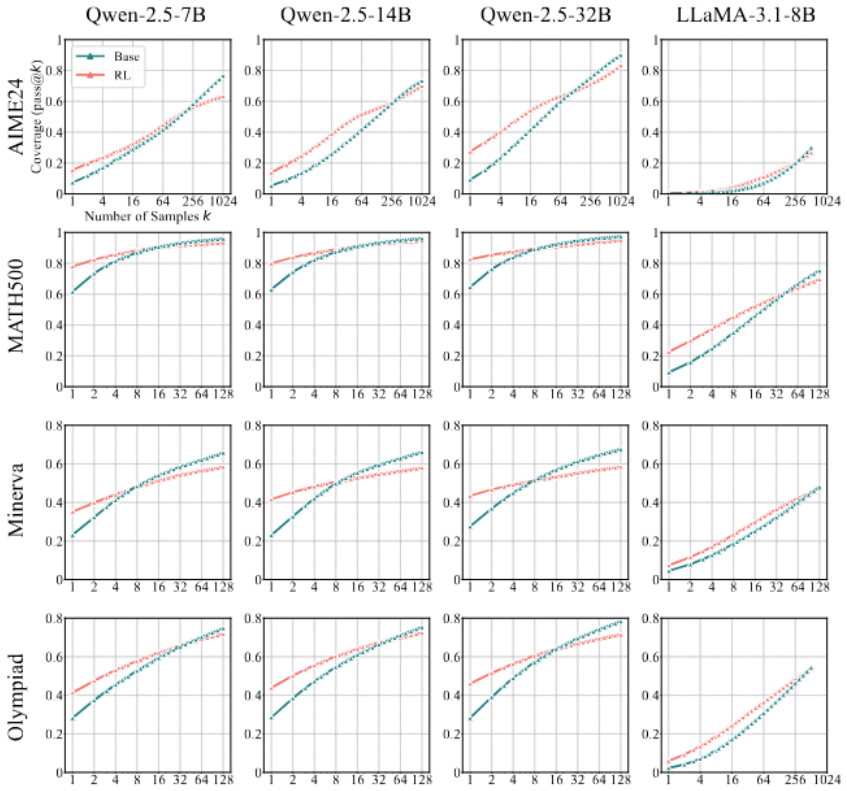


Figure 2: Pass@k curves of base models and their RLVR-trained counterparts across multiple mathematical benchmarks. When k is small, RL-trained models outperform their base versions. However, as k increases to the tens or hundreds, base models consistently catch up and surpass RL-trained models. More results on GSM8K and AMC23 can be found at Figure 9.

Pass@K used by the authors of the paper, instead of more traditional metrics (e.g., Pass@1, Best-of-N, Majority Voting).

Why? What are the authors trying to measure?

RLVR (coding benchmarks)

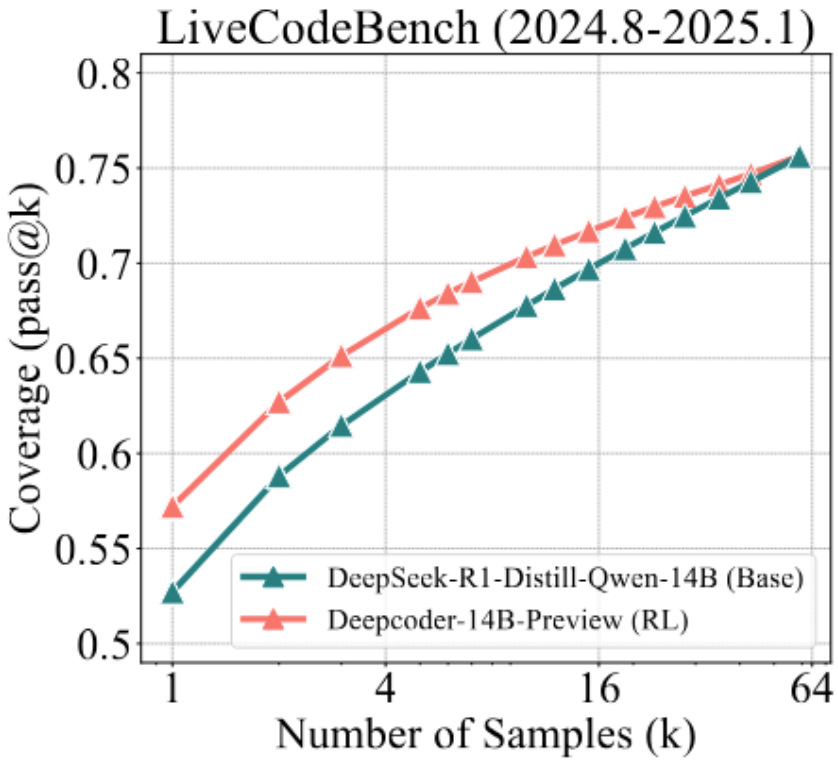


Figure 3: RLVR for Coding.

Author's summary

1. $\cdots$ problems solved by the RLVR model are also solvable by the base model;
   - observed improvement in average scores stems from more efficient sampling on these already solvable problems, rather than learning to solve new problems.
2. $\cdots$ after RLVR training, the model often exhibits narrower reasoning coverage compared to its base model.
3. $\cdots$ all the reasoning paths exploited by the RLVR model are already present in the sampling distribution of the base model.
4. $\cdots$ RLVR does not introduce fundamentally new reasoning capabilities and that the reasoning capacity of the trained model remains bounded by that of its base model.

Distillation versus RLVR

- $\cdots$ training data consist of long CoT reasoning traces generated by the teacher model $\cdots$

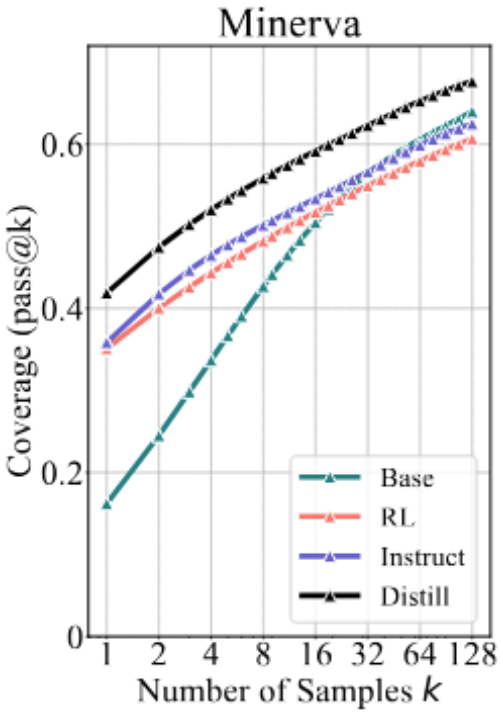


Figure 7: Coverage comparison of base, Instruct, RLVR, and distilled models.

CC's Thoughts

- Interesting paper, experiments and hypothesis are both convincing
- Use of Pass@K instead of other metrics such as Pass@1, Best-of-N, and Majority Voting is interesting for understanding model *capability*
- RLVR is *still* useful; improves performance for Pass@1
- Distillation is potentially more costly, but would has higher *potential* upside (e.g., smaller model size, improved reasoning capabilities)