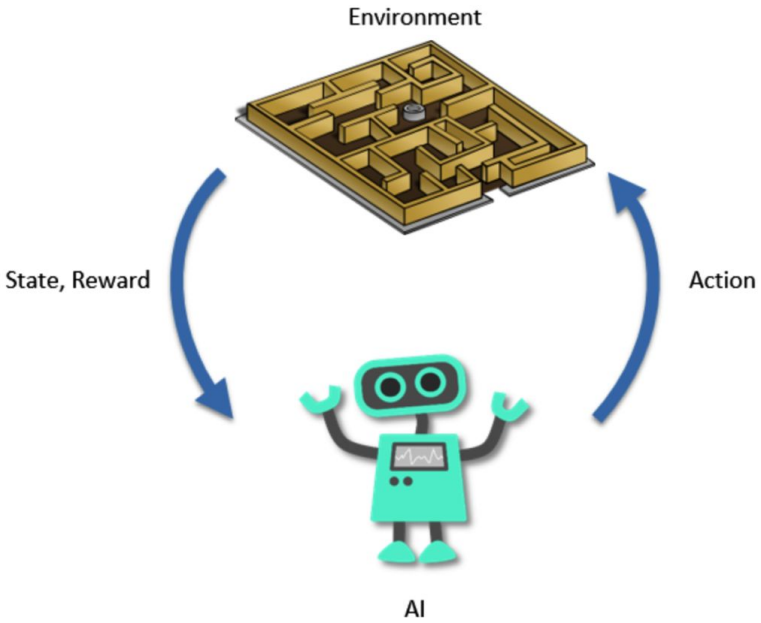
Does RL Instill New Abilities In Models?

9/18/2025

Aidan Alme & Kaavya Chaparala

Why Reinforcement Learning? (Review From Class)

- Adapt model to given task
- Make the foundation model usable without losing grammaticality
- **Expand reasoning capabilities**
- Reinforce desired behavior



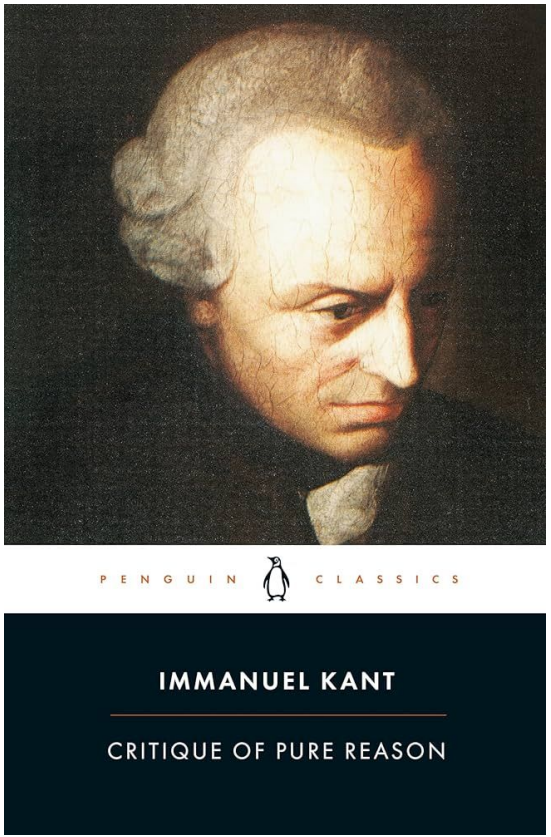
https://upload.wikimedia.org/wikipedia/commons/1/1b/Reinforcement_learning_diagram.svg

Does Reinforcement Learning improve reasoning capabilities beyond the base model?

Yes or no? What do you think?

What is Reasoning?

- Getting things correct?
- **Quid facti vs. quid juris**
- Product of general purpose cognitive abilities
- Verbal, quantitative, spatial
- Applicable to a class or set of problems



https://m.media-amazon.com/images/I/81HFexTeoXL._UF1000,1000_QL80_.jpg

Important Definitions

Action space:

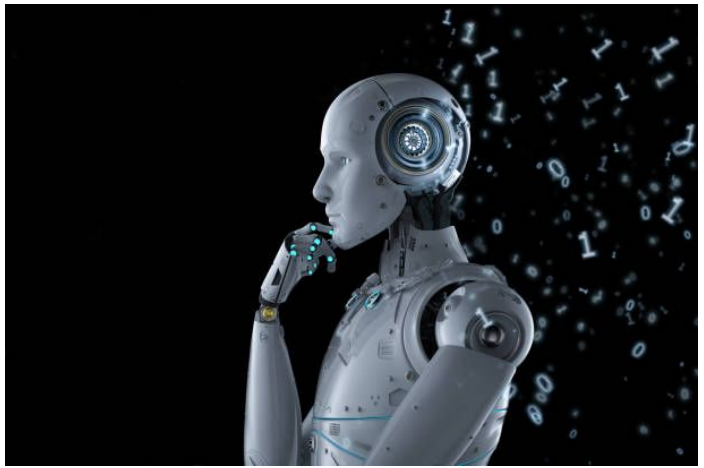
- Set of possible actions agent can take

Language space:

- Embedding space / completions (?)

Chain of thought:

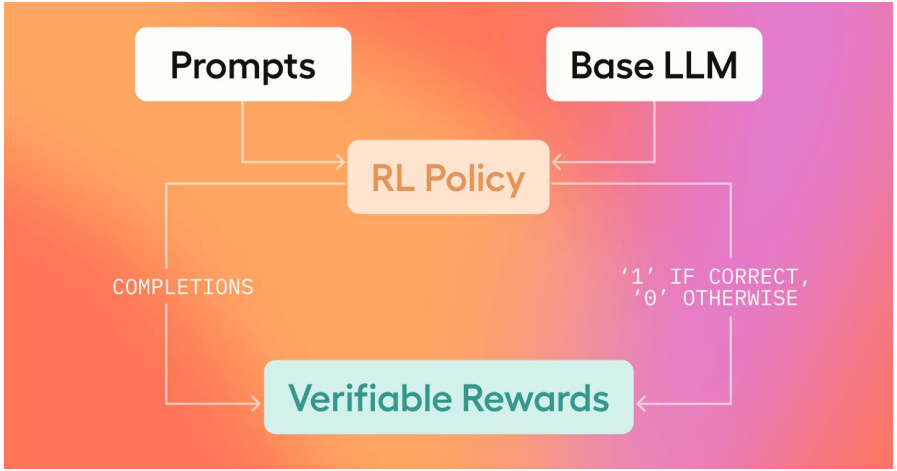
- Prompt-engineering to break down task step-by-step
- Compare: prompt-chaining



https://media.istockphoto.com/id/1096227330/photo/machine-learning-concept.jpg?s=612x612&w=0&k=20&c=bvV5DRZxWVXs80s3gUBCSwWnl8K2DNX1kEKa10P01Gk=

Reinforcement Learning with Verifiable Rewards

- Objective outcome (right or wrong)
- Mathematics, coding, etc.
- Ground truth solution
- Practical and easy to implement
- Tasks involve general reasoning rather than facts or trivia



https://cdn.sanity.io/images/k7elabj6/production/a895a8986032ac0ab2155dacc3a4b7960010c90c-1440x756.png

Important Definitions - pass@k

pass@k is the probability that at least one of the top k outputs generated given a prompt is correct

- Used to assess model reasoning abilities
- pass@1 looks at exact match correctness
- Larger k values examine if the model can generate a wide range of valid outputs

$$\text{pass@}k := \mathop{\mathbb{E}}_{\text{Problems}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right]$$

https://arxiv.org/pdf/2107.03374

Paper & Contributions

ProRL

- Prolonged reinforcement learning
- “The World’s Best 1.5B Reasoning Model”
- *Does RL expand model reasoning capabilities?*

Limit of RLVR

- Analysis of pre-existing methods
- Fine-tuned vs. aligned models
- Pass@k sampling of responses
- *Does RLVR induce novel reasoning?*

ProRL

ProRL - Prolonged Reinforcement Learning Method

- Start with GRPO (Group Relative Policy Optimization)
  - Generate multiple outputs per query
  - Calculate a reward for how well each output answers the query
  - Calculate the advantage for each output
  - *No need for separate ‘Value’ module !*

ProRL - Details

- KL regularization & clip higher
- Reference Policy Reset
- Prolonged RL training periods

ProRL - KL Regularization & higher clipping

Similar to Tuesday's DAPO paper, authors argue that higher clipping encourages exploration

$$\text{clip}(r_\theta(\tau), 1-\epsilon_{low}, 1+\epsilon_{high}) \quad , \epsilon_{high} = 0.4$$

Unlike the DAPO paper, authors keep KL term

- Base model = DeepSeek-R1-Distill-Qwen-1.5B, already well-initialized for CoT

ProRL - Reference Policy Reset & Prolonged Training

Problem: As training progresses, KL penalty will grow and policy updates shrink, limiting the number of training steps that can be taken

Solution: When validation performance plateaus/worsens, reset the reference policy to more recent version of the online policy

Resetting the reference policy enables prolonged training

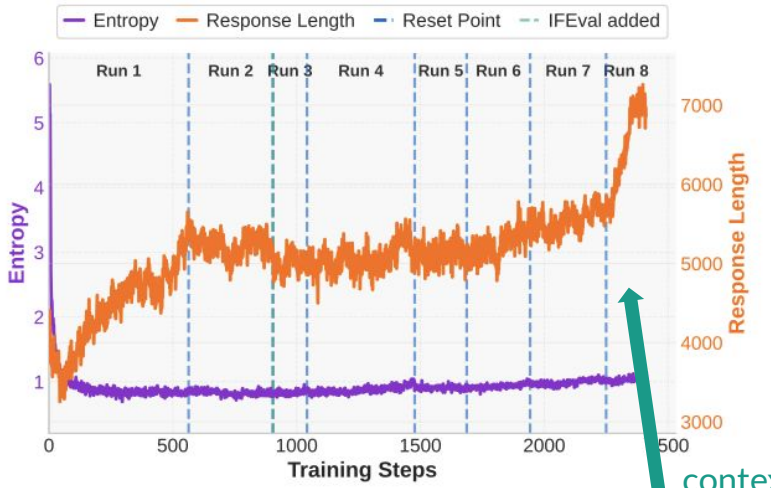
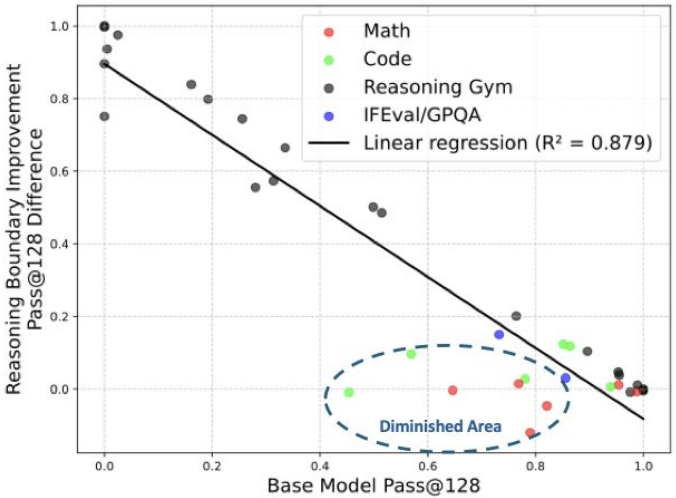


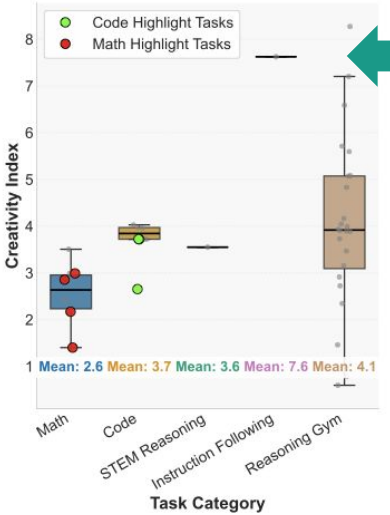
Figure 2: ProRL training dynamics.

ProRL - Results!

ProRL - Performance Compared to Base Model

ProRL expands model's reasoning boundary most on tasks where base model initially struggles.





Tasks with minimal gains post-RL (in the circle) tend to have a lower creativity index, indicating higher overlap with pretraining data.

ProRL - Quantitative Results

Model	Average pass@1
ProRL	60.14
Base Model	63.19

Table 1. Comparison of average pass@1 score across benchmarks for **Math** domain

Model	Average pass@1
ProRL	37.49
Base Model	23.08

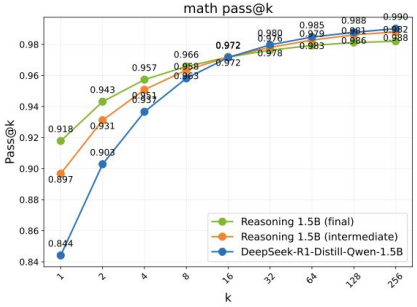
Table 1. Comparison of average pass@1 score across benchmarks for **Code**

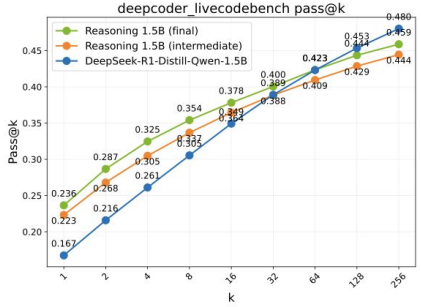
Model	Average pass@1
ProRL	52.29
Base Model	3.49

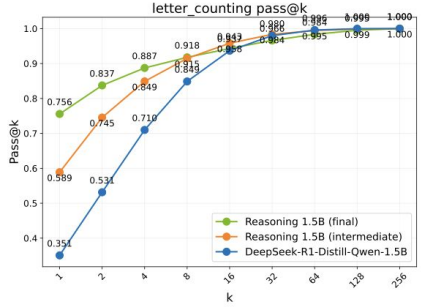
Table 1. Comparison of average pass@1 score across benchmarks for **STEM reasoning**

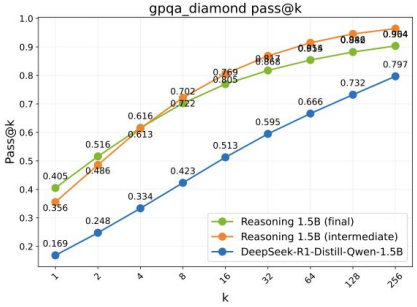
ProRL – Pass@k

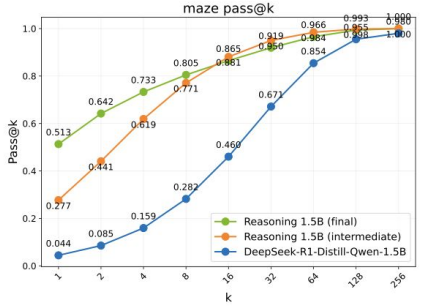
ProRL consistently surpasses base model at big and small k values

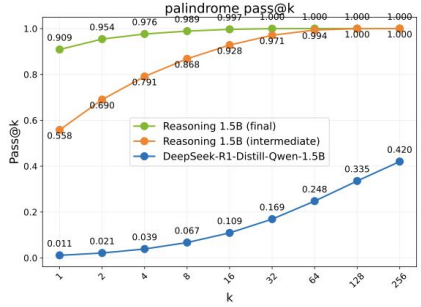


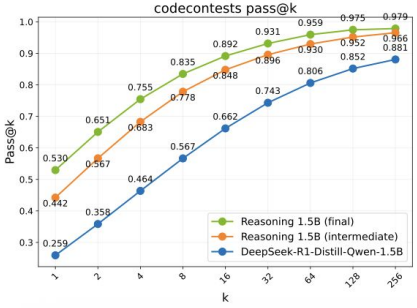


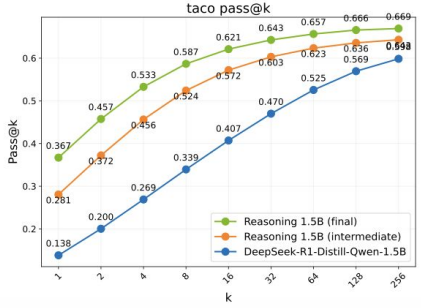


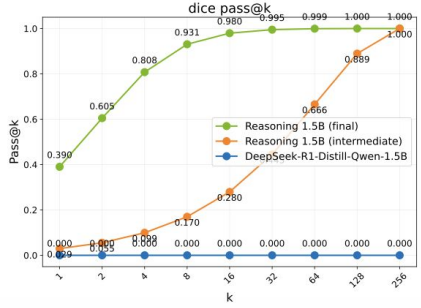












Limits of RLVR

Limit of RLVR - Details

- Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?
- Pass@k sample completions for base vs. aligned & check the proportion correct

Table 1: Experimental setup for assessing RLVR's effect on the reasoning boundaries of LLMs.

Task	Start Model	RL Framework	RL Algorithm(s)	Benchmark(s)
Mathematics	LLaMA-3.1-8B Qwen2.5-7B/14B/32B-Base Qwen2.5-Math-7B	SimpleRLZoo Oat-Zero DAPO	GRPO	GSM8K, MATH500 Minerva, Olympiad AIME24, AMC23
Code Generation	Qwen2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-14B	Code-R1 DeepCoder	GRPO	LiveCodeBench HumanEval+
Visual Reasoning	Qwen2.5-VL-7B	EasyR1	GRPO	MathVista MathVision
Deep Analysis	Qwen2.5-7B-Base Qwen2.5-7B-Instruct DeepSeek-R1-Distill-Qwen-7B	VeRL	PPO, GRPO Reinforce++ RLOO, ReMax, DAPO	Omni-Math-Rule MATH500

The proportion of correct samples from unaligned models surpasses the proportion from aligned ones for high enough k!

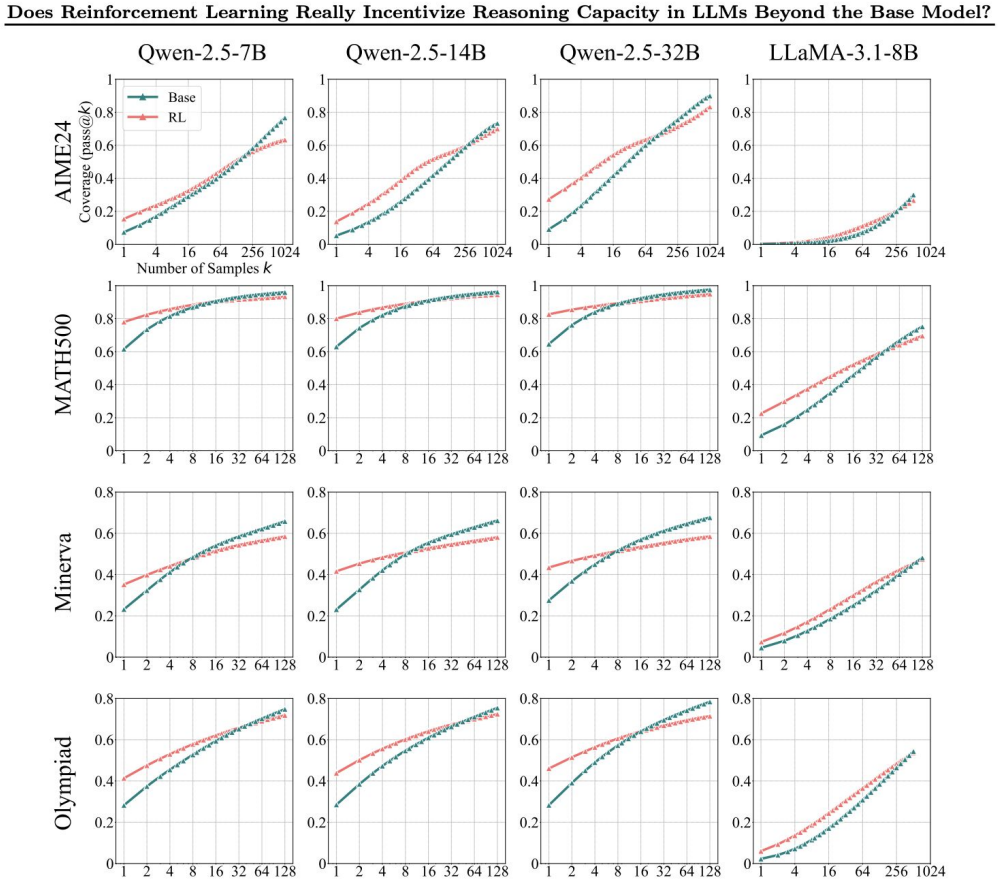


Figure 2: Pass@k curves of base models and their RLVR-trained counterparts across multiple mathematical benchmarks. When k is small, RL-trained models outperform their base versions. However, as k increases to the tens or hundreds, base models consistently catch up and surpass RL-trained models. More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

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

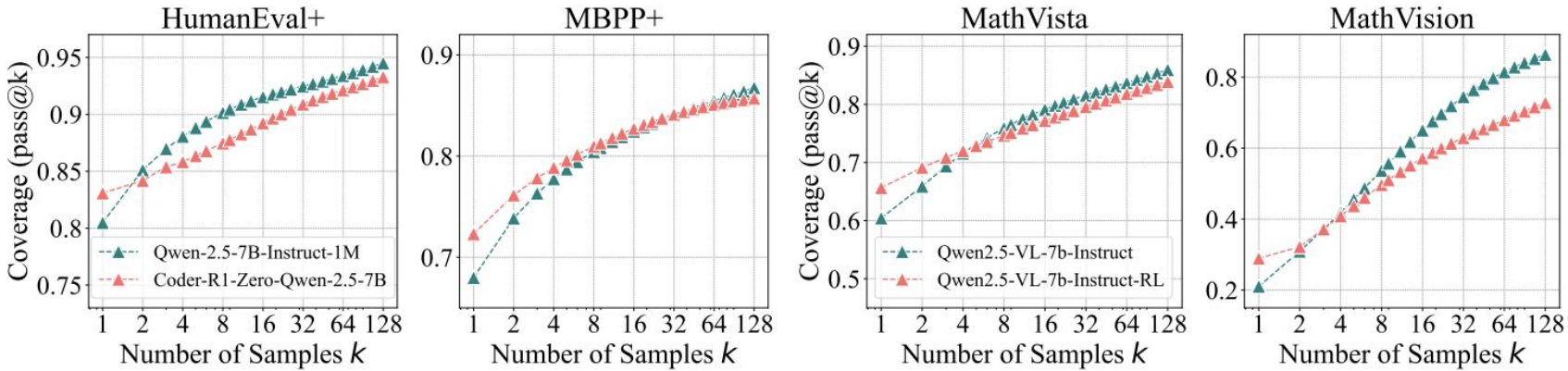


Figure 4: Pass@k curves of base and RLVR models. **(Left)** Code Generation. **(Right)** Visual Reasoning.

Closer look at the same phenomenon from last slide.

Notice the proximity in accuracy.

Pruning or reasoning?

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

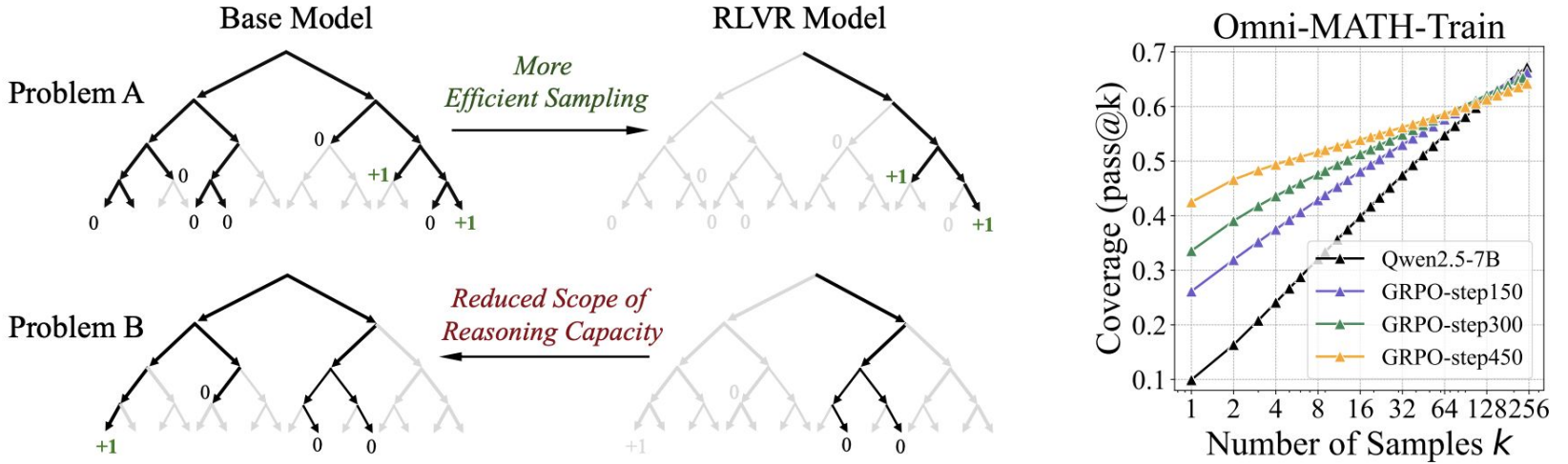


Figure 1: **(Left)** The effect of current RLVR on LLM's reasoning ability. Search trees are generated by repeated sampling from the base and RLVR-trained models for a given problem. Grey indicates paths that are unlikely to be sampled by the model, while **black** indicates paths that are likely to be sampled. Green indicates correct paths, which has positive rewards. Our key finding is that all reasoning paths in the RLVR model are already present in the base model. For certain problems like Problem A, RLVR training biases the distribution toward rewarded paths, improving sampling efficiency. However, this comes at the cost of reduced scope of reasoning capacity: For other problems like Problem B, the base model contains the correct path, whereas that of the RLVR model does not. **(Right)** As RLVR training progresses, the average performance (*i.e.*, pass@1) improves, but the coverage of solvable problems (*i.e.*, pass@256) decreases, indicating a reduction in LLM's reasoning boundary.

Limit of RLVR - Details

- Narrower reasoning - maximization of expectation?
- Reasoning paths exist in base model - superset
    - Cannot solve new problems
- Minor variation in RL / alignment algorithms
- Distillation does cause improvement
- Vast action space
- Similar perplexity

https://www.vintagetreecare.com/wp-content/uploads/2014/01/tree-pruning.jpg

https://miro.medium.com/1*3DvSjCLORuexj3Y-G-r8hw.jpeg

Conclusions

ProRL

“The weaker the start the stronger the gain”

- ProRL model is poorer on tasks where the base model already performed well
- stronger on tasks where the tasks where base model was weak

Limit of RLVR

Capacity already present in base model

- No new reasoning
- Effective sampling
- Reduced reasoning?

Does RL *really* instill reasoning capabilities?

Contradictory Results?