Rethinking LLM Reasoning

Xiang Yue

Carnegie Mellon University

xyue2@andrew.cmu.edu

https://xiangyue9607.github.io

Carnegie Mellon University

Language Technologies Institute

Solving Complex Reasoning Problems with LLMs

Google A.I. System Wins Gold Medal in International Math Olympiad

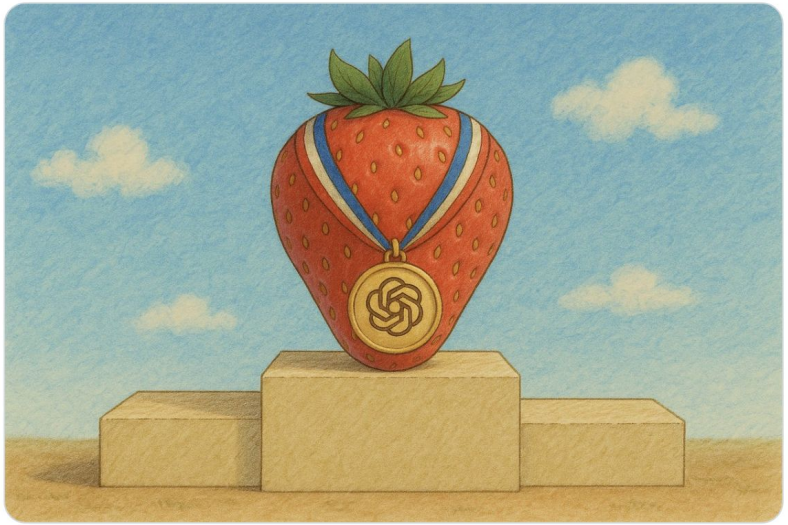
OpenAI said it, too, had built a system that achieved similar results.

Listen to this article · 4:20 min Learn more

Share full article

Alexander Wei ✓
@alexwei_

1/N I'm excited to share that our latest @OpenAI experimental reasoning LLM has achieved a longstanding grand challenge in AI: gold medal-level performance on the world's most prestigious math competition—the International Math Olympiad (IMO).



12:50 AM · Jul 19, 2025 · **5.1M** Views

AI Reasoning Has a Long History

Allen Newell, Herbert Simon who created "Logic Theorist," 1st thinking machine in 1955

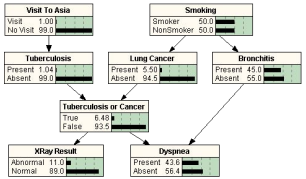
Prolog is a logic programming language that has automated theorem proving and computational linguistics

Sample Prolog Program:
`grandmother.pl`

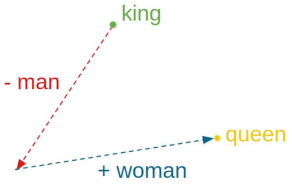
```
grandmother(X, Y) :- mother(X, Z), parent(Z, Y).
parent(X, Y) :- mother(X, Y).
parent(X, Y) :- father(X, Y).

mother(mary, stan).
mother(gwen, alice).
mother(valery, gwen).
father(stan, alice).
```

Bayesian networks and probabilistic models handle uncertainty and make informed decisions.



Neural networks advance semantic understanding and relational reasoning.



LLMs revolutionize natural language understanding and reasoning.

Foundational Concepts and Rule-based Systems (1950s)

Knowledge Representation & Symbolic AI (1970s-1990s)

Statistical and Probabilistic Methods (2000s)

Neural Reasoning (2010s)

LLM Reasoning (2020s)

Why LLM Reasoning? What is Different Today?

Why Reasoning with LLMs?

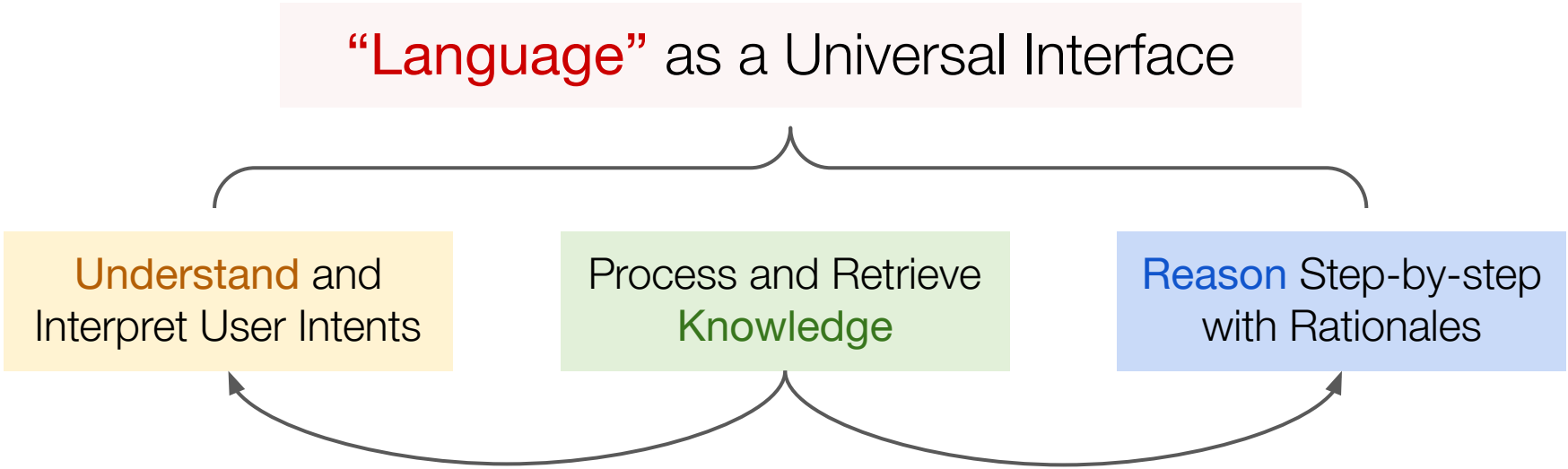
"Language" as a Universal Interface

Understand and Interpret User Intents

Process and Retrieve Knowledge

Reason Step-by-step with Rationales

Why Reasoning with LLMs?



Knowledge as the foundation for understanding and reasoning

This enables models to generalize and reason over ***unseen*** questions

Why we care about reasoning?

"**Reason** is the capacity of *applying logic by drawing valid conclusions* from new or existing information… , and is normally considered to be **a distinguishing ability possessed by humans**"

"**Intelligence** has been defined in many ways: the capacity for abstraction, **logic**, **understanding**, self-awareness, learning, emotional knowledge, **reasoning**, **planning**, creativity, critical thinking, and **problem-solving"**

What is Intelligence?

Intelligence as a *collection of task-specific skills*

"Much of the human cognitive function is the result of special-purpose adaptations to solve specific problems." --Charles Darwin

"AI is the science of making machines capable of performing tasks that would require intelligence if done by humans."
--Marvin Minsky

Intelligence as a *general learning ability*

"Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets." --Alan Turing

"AI is the science and engineering of making machines do tasks they have never seen"
--John McCarthy

Intelligence measures a model's ability to *efficiently acquire and apply skills to achieve goals* in *novel and dynamic* environments

(My view on "Intelligence")

Towards that, **“Generalizing to Novel and Unseen Tasks”** is the key

"Why are we keeping pushing math reasoning?"

- Math reasoning is clearly defined and easy to verify.
- AI / CS people understand math well.
- Math could be a proxy of general reasoning. Improving math reasoning could transfer to general LLM capability.

"Why are we keeping pushing math reasoning?"

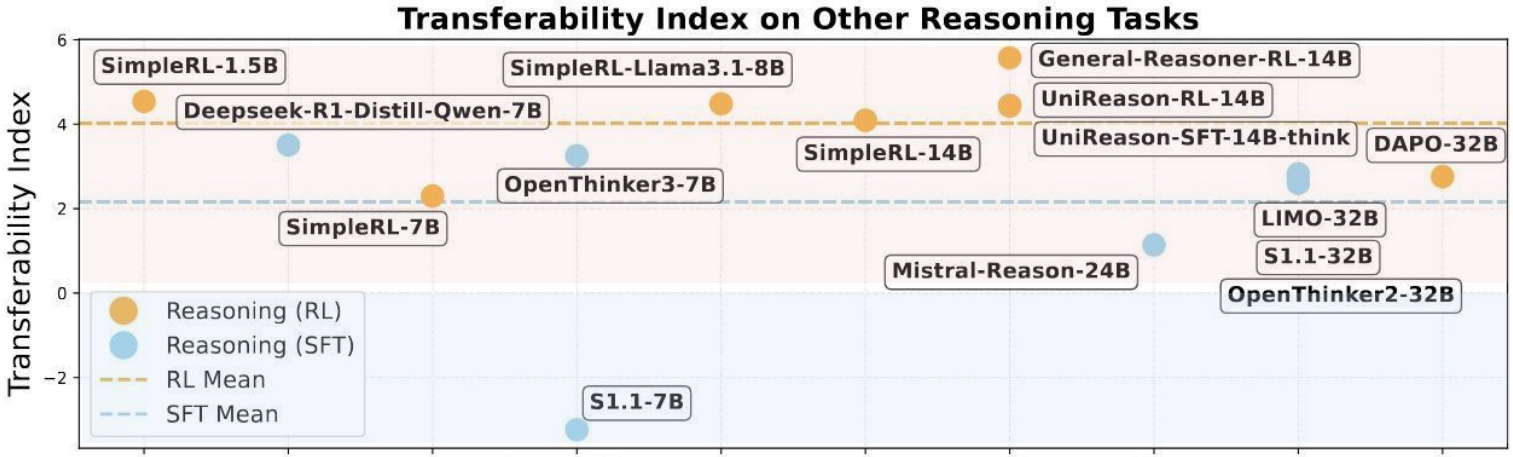
- Math reasoning is clearly defined and easy to verify.
- AI / CS people understand math well.
- Math could be a proxy of general reasoning. Improving math reasoning could transfer to general LLM capability.

That is our hope. But is it true?

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

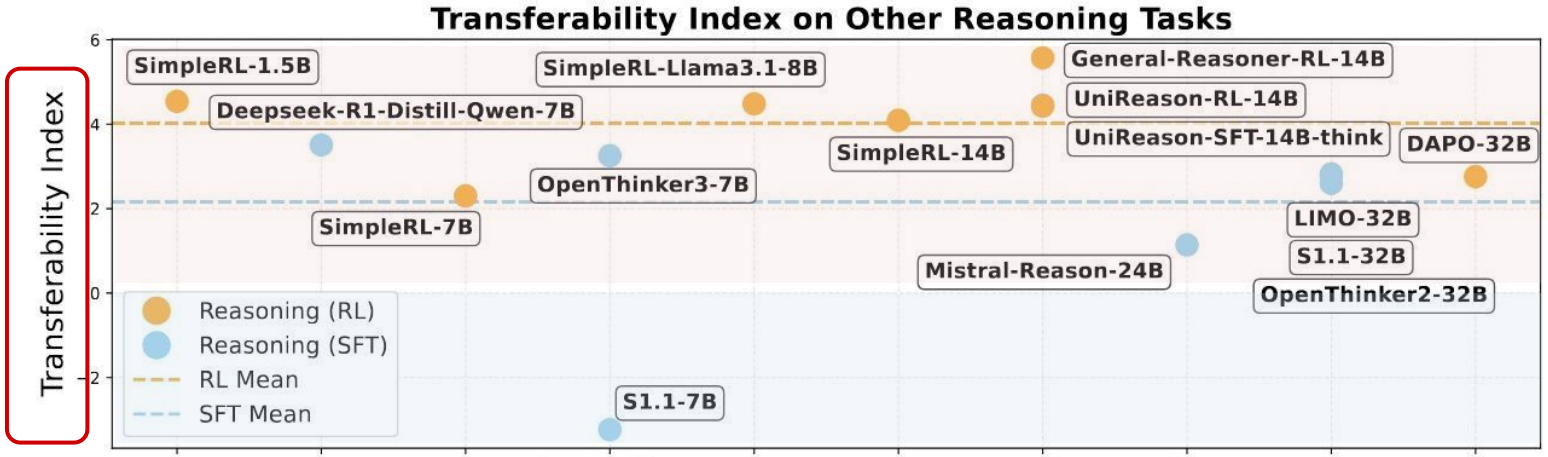
Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, **Xiang Yue**

How math reasoning transfers to **other reasoning** tasks?



Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

How math reasoning transfers to **other reasoning** tasks?



TI measures the performance delta ratio (base->fine-tuned)

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Transferability Index (TI)

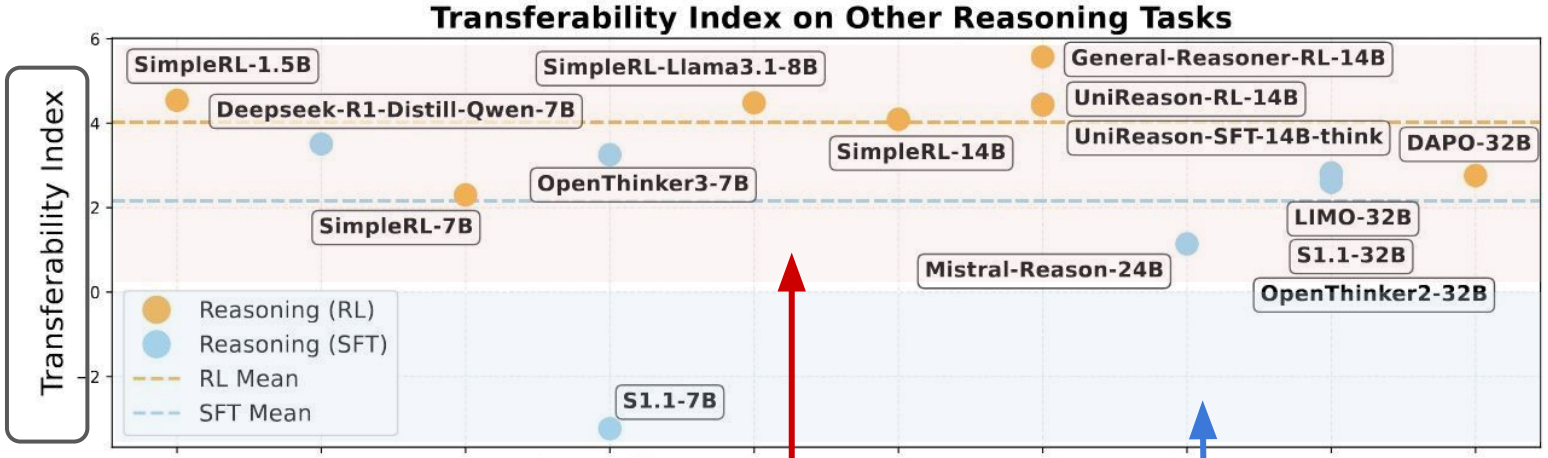
Let $\mathcal{B}_g$ be the set of benchmarks in group $g \in \{\texttt{math}, \texttt{other}, \texttt{non}\}$, corresponding to each of our task groups: math reasoning, other reasoning and non-reasoning. Let $|\mathcal{B}_g|$ be its cardinality. For each benchmark $b \in \mathcal{B}_g$ we have scores R_b^{model} and R_b^{base}. We define the group-level relative gain as the average of per-benchmark gains:

$$\Delta R_g = \frac{1}{|\mathcal{B}_g|} \sum_{b \in \mathcal{B}_g} \frac{R_b^{\text{model}} - R_b^{\text{base}}}{R_b^{\text{base}}}, \quad g \in \{\texttt{math}, \texttt{other}, \texttt{non}\}.$$

Next, the two Transferability Indices are

$$\text{TI}_{\text{other}}(\%) = \frac{\Delta R_{\text{other}}}{\Delta R_{\text{math}}} \times 100, \qquad \text{TI}_{\text{non}}(\%) = \frac{\Delta R_{\text{non}}}{\Delta R_{\text{math}}} \times 100.$$

How math reasoning transfers to **other reasoning** tasks?



TI measures the performance delta ratio (base->fine-tuned)

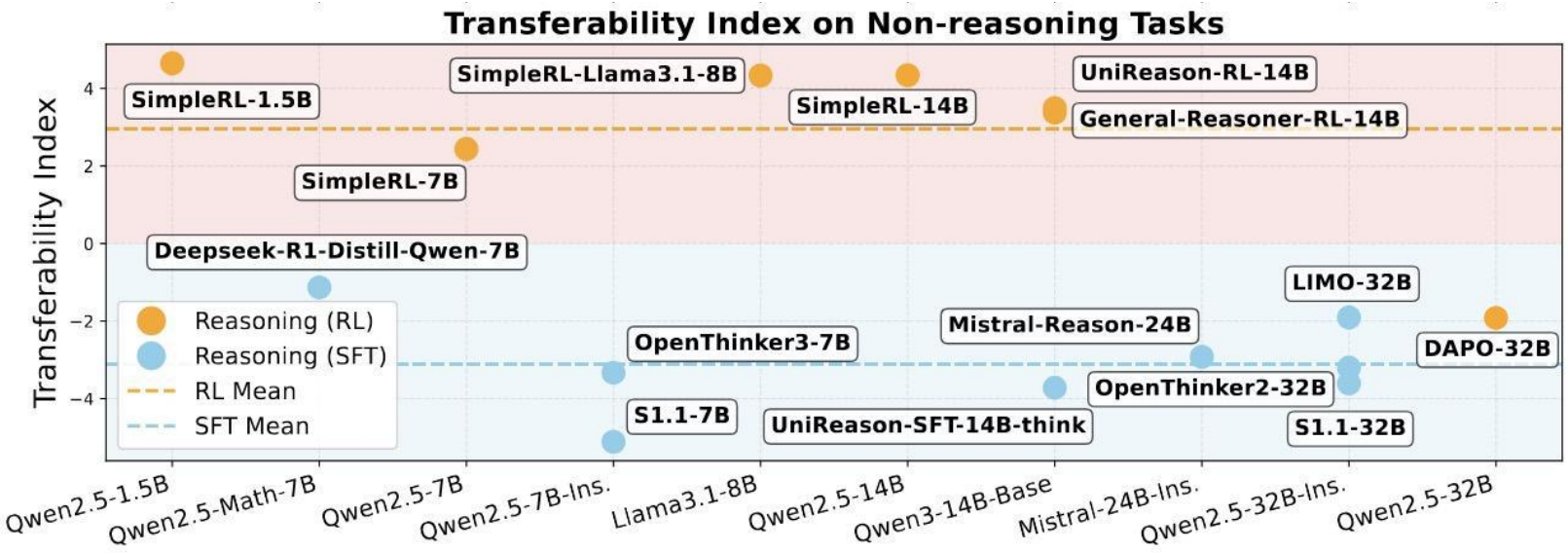
TI > 0: improved math could transfer to other domains

TI < 0: improved math could not transfer to other domains

Improved math reasoning could transfer to other reasoning tasks like coding and science reasoning

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

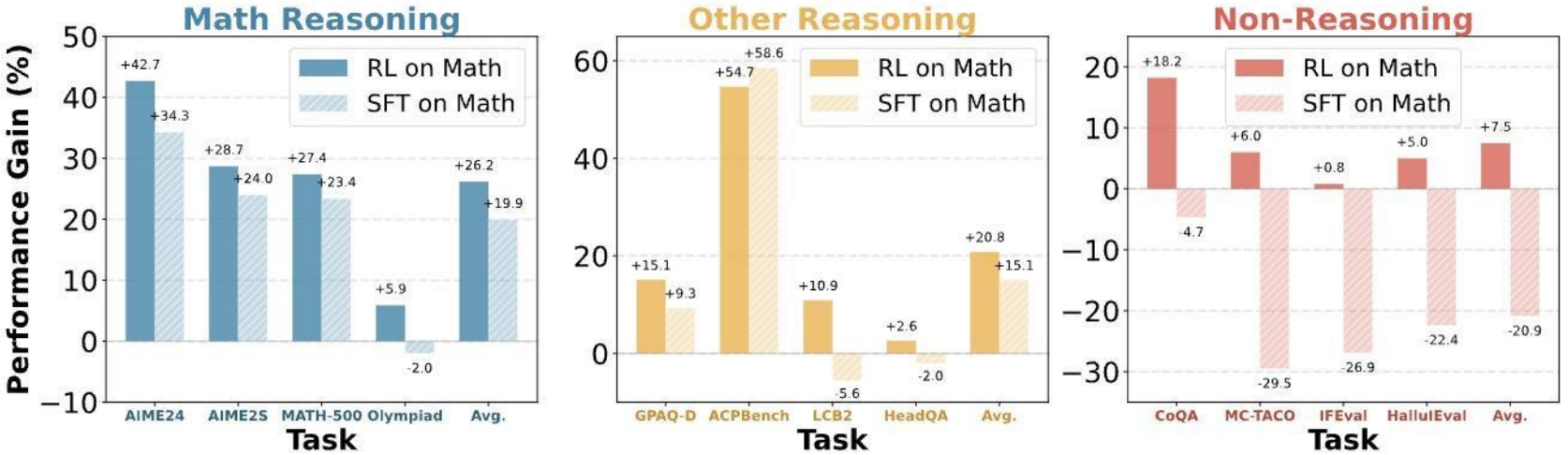
How math reasoning transfers to **non-reasoning** tasks?



Improved math reasoning could transfer to **non-reasoning** tasks mostly when models are **trained with RL**

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Controlled Experiments

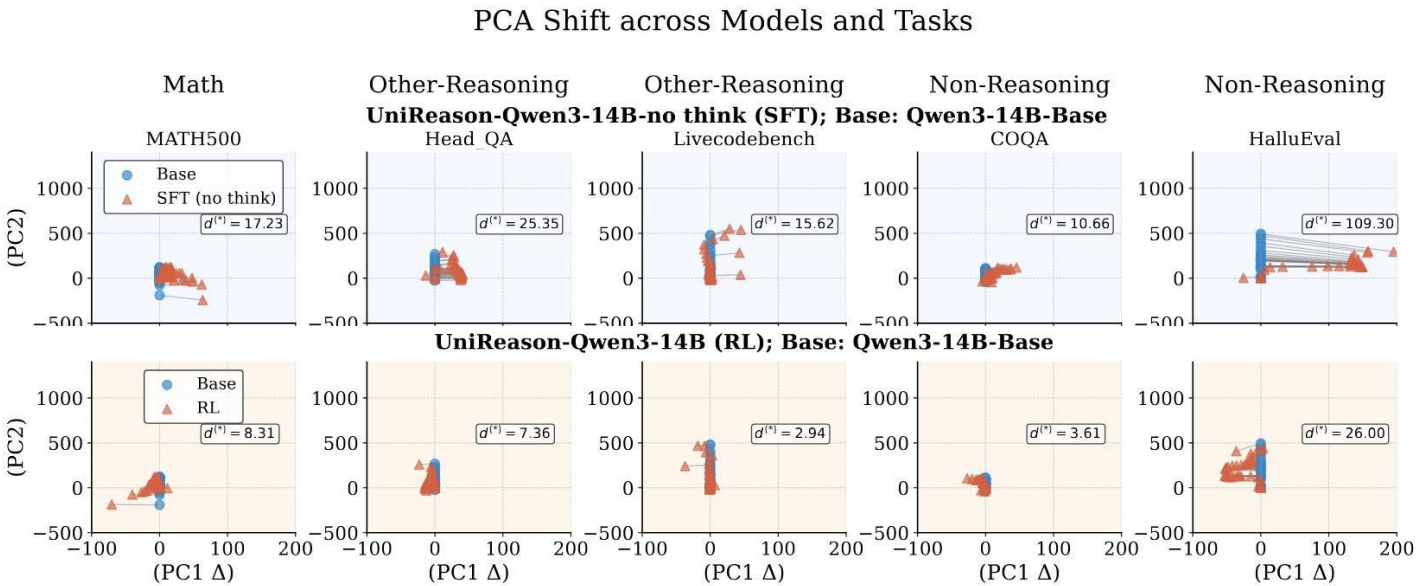


- While SFT-trained models partially generalize to other reasoning tasks, they show limited transfer to non-reasoning tasks.
- In contrast, RL-trained models exhibit broader generalization across both reasoning and non-reasoning scenarios.

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Why RL can lead to more generalization?

RL exhibits minor distribution shifts than base model

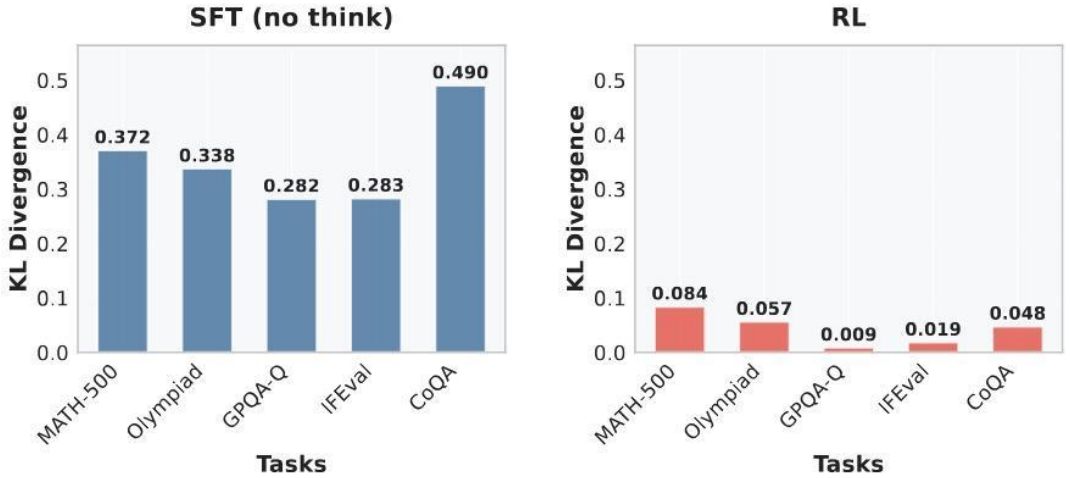


Hidden Representation Level:

- We employ PCA analysis to examine the internal hidden state of SFT and RL model.
- $d\,(*)$ is the Euclidean distance between representation centroids before and after training.

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

RL exhibits minor distribution shifts than base model

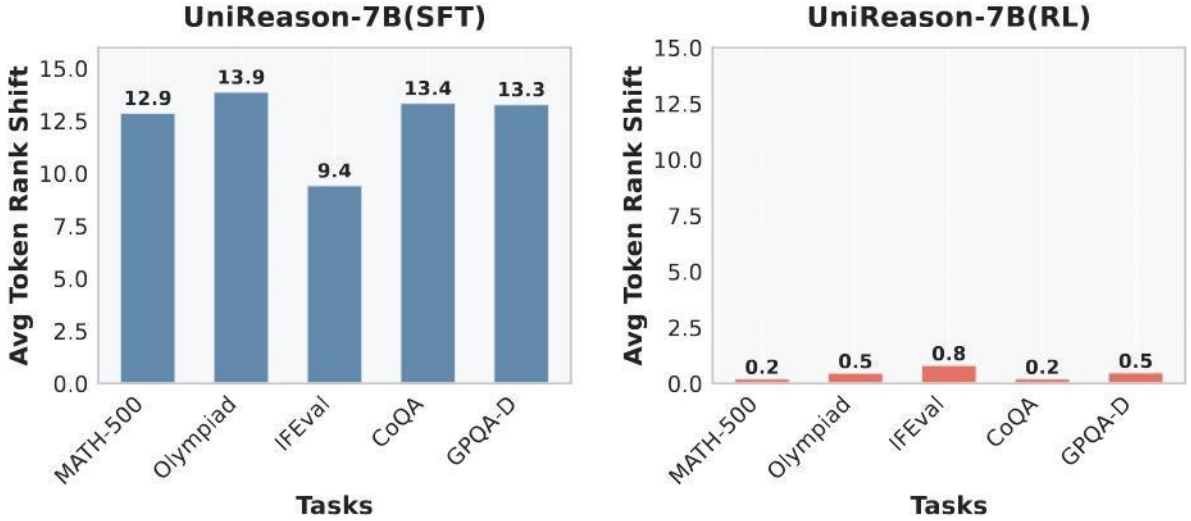


Token Distribution Level:

- KL divergence analysis of RL and SFT models.
- Higher KL divergence indicates greater distribution shifts from the original backbone model.
- We observe that RL models consistently exhibit significantly lower KL divergence compared to SFT models across different tasks, suggesting less distribution shift during training..

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

RL exhibits minor distribution shifts than base model

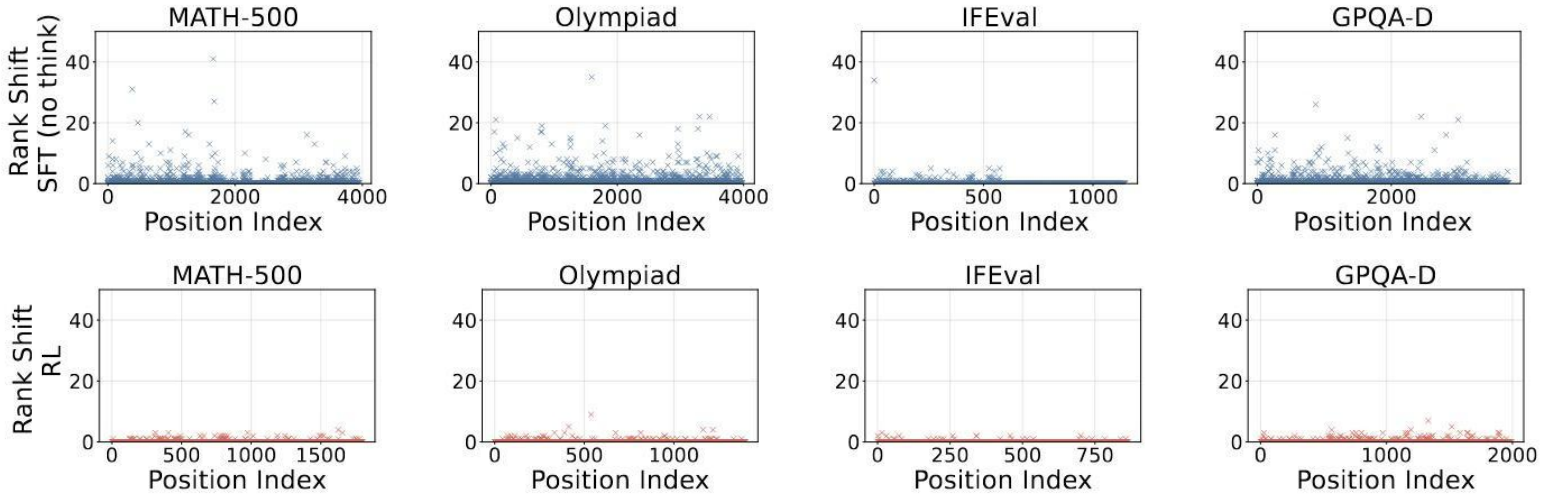


Token Distribution Level:

- Average token rank shift of SFT and RL models compared to their base models.
- We generate tokens using fine-tuned models and evaluate their rank shifts under the base model's distribution.
- RL only shifts no less than 0.5 ranks on average compared with the base model.

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

RL exhibits minor distribution shifts than base model

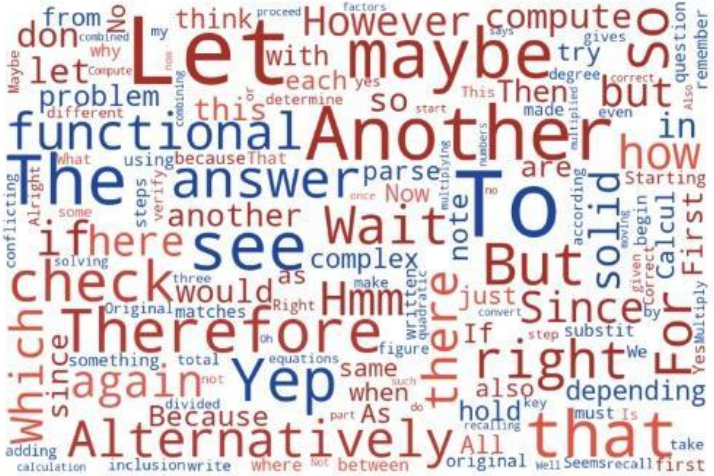


Token Distribution Level:

- Visualization of token rank shifts across different position indices for both reasoning and non-reasoning tasks.
- We observe that RL models exhibit less token rank shifts while SFT models demonstrate substantial rank shifts across numerous positions throughout the sequence..

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Visualization of Top Shifted Tokens



Token Distribution Level:

- Tokens are extracted based on frequency and rank shifts, then categorized as logical-structural words or content-specific words.
- After training on MATH data, RL model shifts logic-related tokens such as But and So, while the SFT model shifts various types of tokens, including many irrelevant noisy tokens to the task.

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Case Study

Domain	Query	Model	Shifted Tokens
Reasoning Task	*Ten treeks weigh as much as three squigs and one goolee. Two treeks and one goolee are equal in weight to one squig. How many treeks' weight equals one squig?*	**RL Model**	**(Only 15 tokens experienced rank shift when decoded in the base model)** `In a Now Now define for number second 2 Now , add This`
		SFT Model	**(390 tokens experienced rank shift)** `they The again conflicting but m Alternatively make have Hmm hold equations Wait For find check Let maybe using written Original pl contrad So There Wait solve I 's Alternatively Alright so First solving a either check conflicting write Correct here another Like where ? Still From where The question / . The where here where equations Therefore problem check if was the ? equations together . answer I For or For Wait matrices this about m either and solve combined 1 problem ten Let . equation That If...`

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Case Study

		RL Model	**(Only 14 tokens experienced rank shift when decoded in the base model)** `Write « but » Res formally much step grown will once Full`
Non-reasoning Task	*Write an email to my boss telling him that I am quitting. The email must contain a title wrapped in double angular brackets*	**SFT Model**	**(158 tokens experienced rank shift)** `Hmm Alright Wait Wait Wait try Another Maybe Another Alternatively Wait but Wait Wait Diamond On A check Who Starting user generate original (original example make structure So follow The instructions user Let (First (check says doesn . to But willingness generated 's : but says wants so has follow . . The structure the the first is But is structured with However who step like given repeated then also mention answer adding Let the . concise Since like straightforward . effective maybe wants But particular The answer the answer that would « The which original instruction which with the)". the first context . the providing Email of The The I first exactly then provide ...`

- RL models selectively shift task-relevant or logic-token tokens
- In contrast, SFT models inappropriately introduce reasoning-related tokens in non-reasoning queries, leading to unnecessary overthinking that detracts from performance.

Which RL component matters for generalization?

What is the fundamental difference of SFT and RL?

A unified loss of SFT and RL (all likelihood)

For prompts x and completions y, let $\pi_\theta(y \mid x)$ denote the current policy, and $\pi_{\text{ref}}(y \mid x)$ denote a fixed reference policy (e.g., the initialization).

Supervised fine-tuning (SFT). With reference completions $y^\star$, the objective is:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y^\star)\sim\mathcal{D}}\left[\log \pi_\theta(y^\star \mid x)\right]. \tag{1}$$

Reinforcement Learning (RL). For the same prompts, we sample $y \sim \pi_\theta$ and weight each sample by an *advantage* $A(x, y)$:

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[A(x,y)\cdot \log \pi_\theta(y \mid x)\right]. \tag{2}$$

Usually, a KL term is added to prevent the policy model from being too far away from the initialized model. We generalize these objectives using:

$$\mathcal{L}_{q,w,\beta}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim q(\cdot|x)}\left[w(x,y)\cdot \log \pi_\theta(y \mid x)\right] + \beta\, \mathbb{E}_{x\sim\mathcal{D}}\left[\text{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x))\right]. \tag{3}$$

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Which Components of RL Drive Generalization?

There are three major factors in this surrogate loss:

- a *sampling distribution* $q(y \mid x)$ (whether being on-policy or off-policy),
- *credit weights* $w(x, y)$ (uniformly one or weighted by advantage),
- a *KL regularization weight* $\beta \geqslant 0$ against a reference policy π_{ref}.

Setting	**Sampling** q	**Weights** w	**KL Reg.** β
Off-policy SFT	$\delta_{y=y^\star}$	1	0
On-policy SFT	π_θ	1 (reject sample)	0
Off-policy RL	$\delta_{y=y^\star}$	Advantage A_t	0
On-policy RL (no KL)	π_θ	Advantage A_t	0
On-policy RL	π_θ	Advantage A_t	> 0

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

Which Components of RL Drive Generalization?

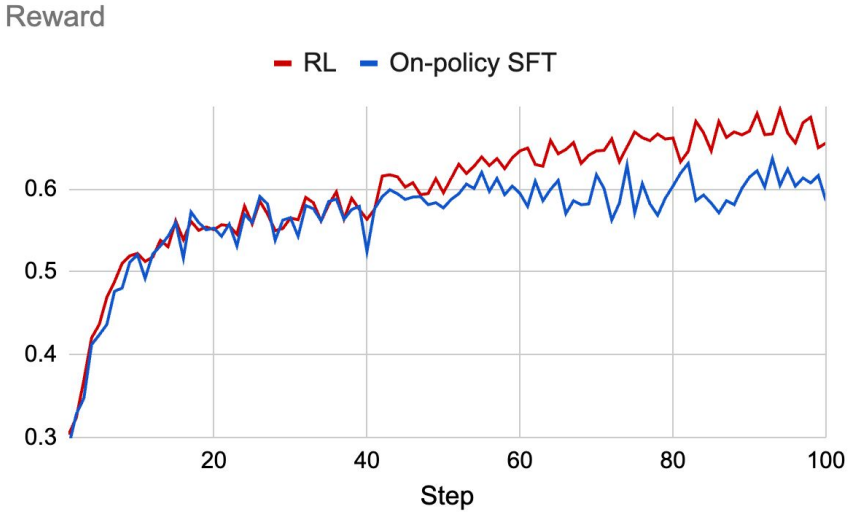
Setting	Sampling q	Weights w	KL Reg. β
Off-policy SFT	$\delta_{y=y^\star}$	1	0
On-policy SFT	π_θ	1 (reject sample)	0
Off-policy RL	$\delta_{y=y^\star}$	Advantage A_t	0
On-policy RL (no KL)	π_θ	Advantage A_t	0
On-policy RL	π_θ	Advantage A_t	> 0

We found:

1. Sampling Distribution (on-policy) matters. SFT with on-policy sampling can transfer the reasoning capability as well.
2. Negative gradient enables longer chains and more robust and improved performance
3. KL penalty has minor impacts for RL (minor impact on all the performances)

Maggie Huan*, Yuetai Li*, Tuney Zheng*, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and **Xiang Yue**. "Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning." *arXiv 2025* ***(*: my advisee)***

RL vs On-policy SFT



Negative gradient enables longer chains and higher performance ceiling

=> It seems like that RL exactly demonstrates transferability and mitigates forgetting during post-training.
However, does RL really forget nothing?

=> We show that in the following paper:

- RL does not exhibit overall level forgetting

(no overall performance degradation)

- **But RL still experiences individual level forgetting!**

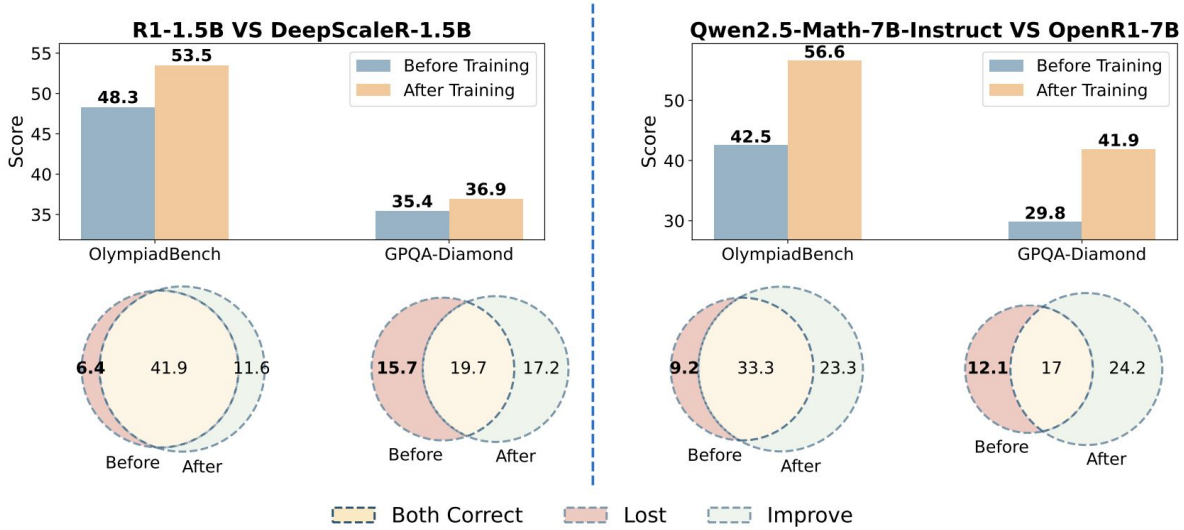
Temporal Sampling for Forgotten Reasoning in LLMs

Yuetai Li,[1] Zhangchen Xu,[1] Fengqing Jiang,[1] Bhaskar Ramasubramanian[3]
Luyao Niu,[1] Bill Yuchen Lin,[1] and **Xiang Yue**,[2] Radha Poovendran[1]

1. University of Washington, 2. Carnegie Mellon University
3. Western Washington University

Overall Score Cannot Tell Everything

Despite the improvement of overall performance, a considerable percentage of questions (from 6.1% to 16%) answered correctly by the base model may be answered incorrectly after RL/SFT.

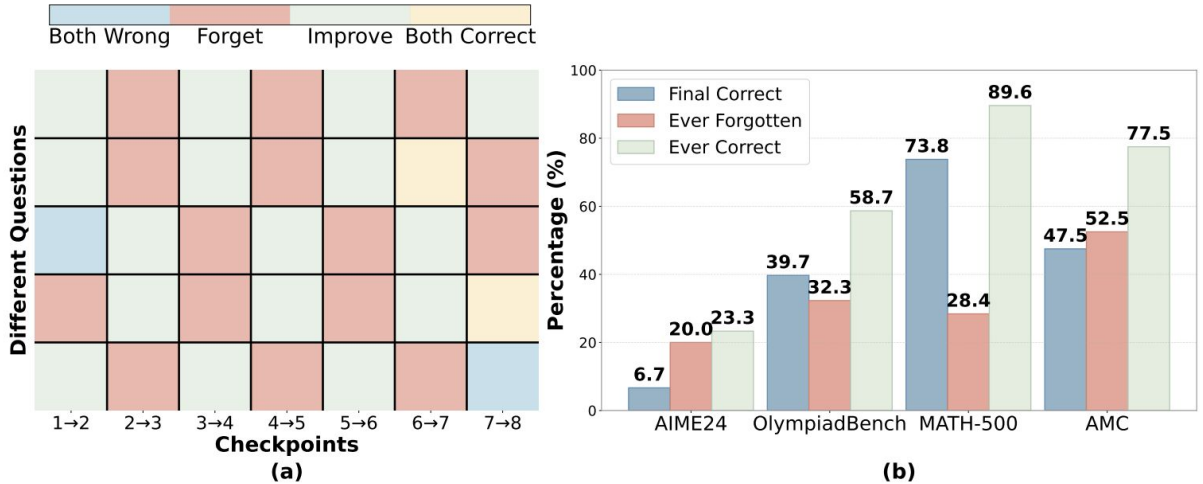


Fine-tuned models like DeepscaleR-1.5B and OpenR1-7B outperform the base model overall but also forget many questions the base model answered correctly.

Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Temporal Forgetting

Benchmark questions may oscillate between correct and incorrect states across checkpoints during training. A considerable percentage of questions (from 6.4% to 56.1%) are answered correctly at least once by some checkpoint during training but are ultimately incorrect in the final checkpoint.

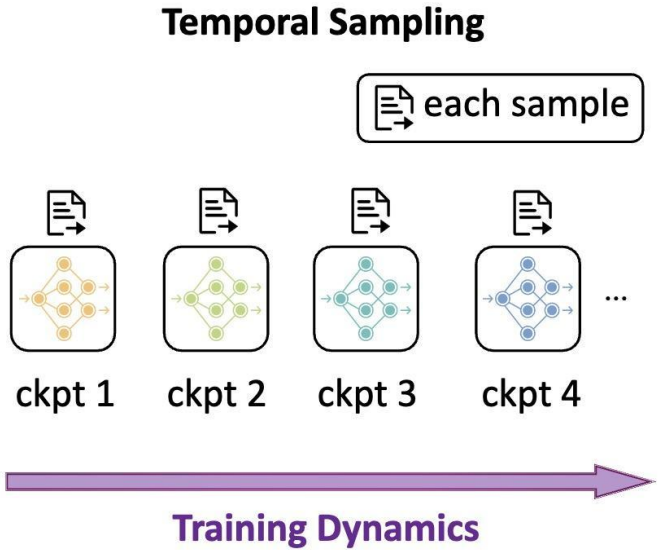


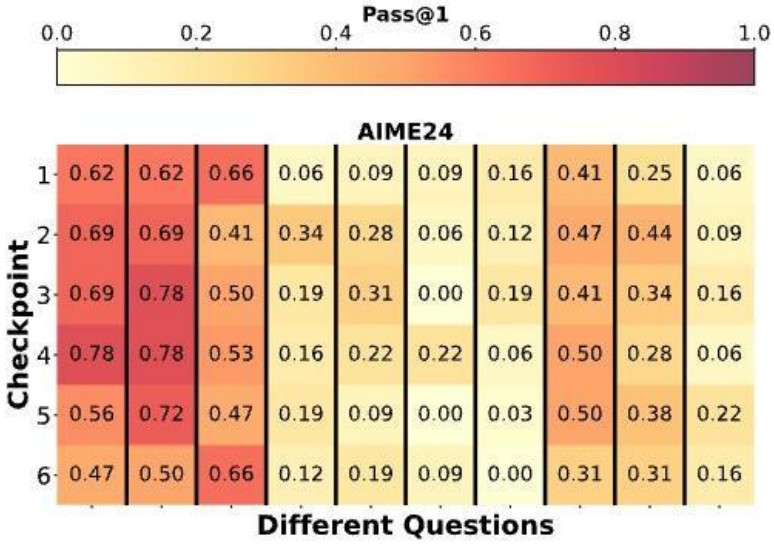
(a) Answer correctness trajectories for different questions across training checkpoints, illustrating solutions oscillate between correct and incorrect states.
(b) Percentage of questions that are ever forgotten or ever correct at some checkpoint during GRPO.

Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Temporal Sampling

- (Left) We utilizes training dynamics as a source of answer diversity by distributing inference samples across multiple distinct checkpoints from the training trajectory, rather than relying solely on the final checkpoint.
- (Right) Pass rate distribution across different training checkpoints when evaluated on AIME24. Individual problems show varying pass rates over time.

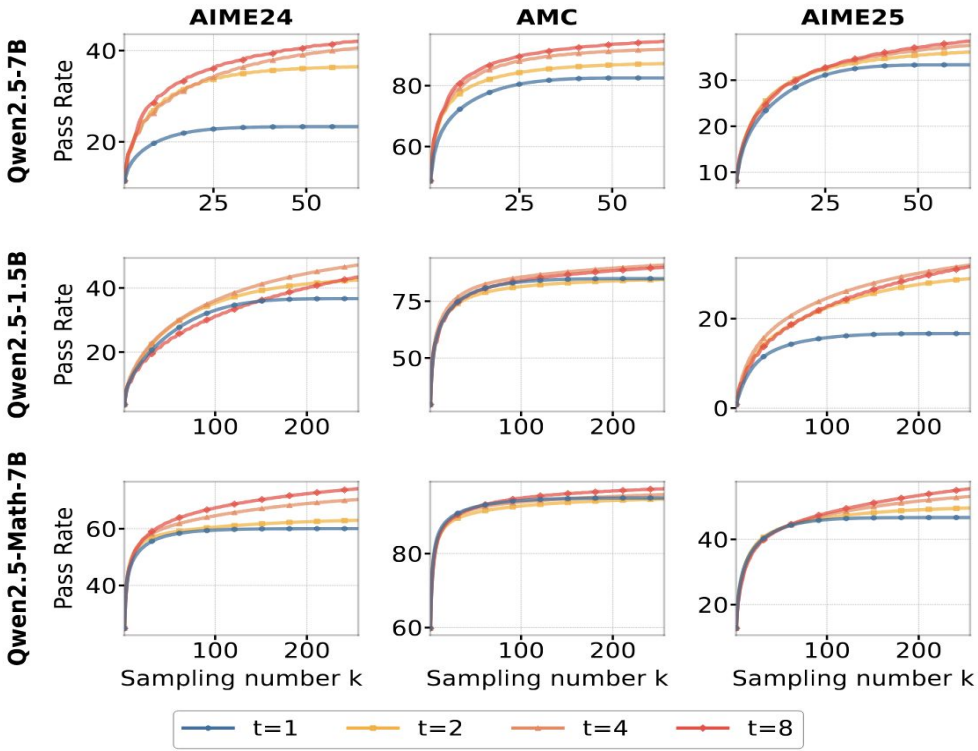




Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Temporal Sampling

Temporal Sampling has better test-time scaling performance than sampling only on the final checkpoint.

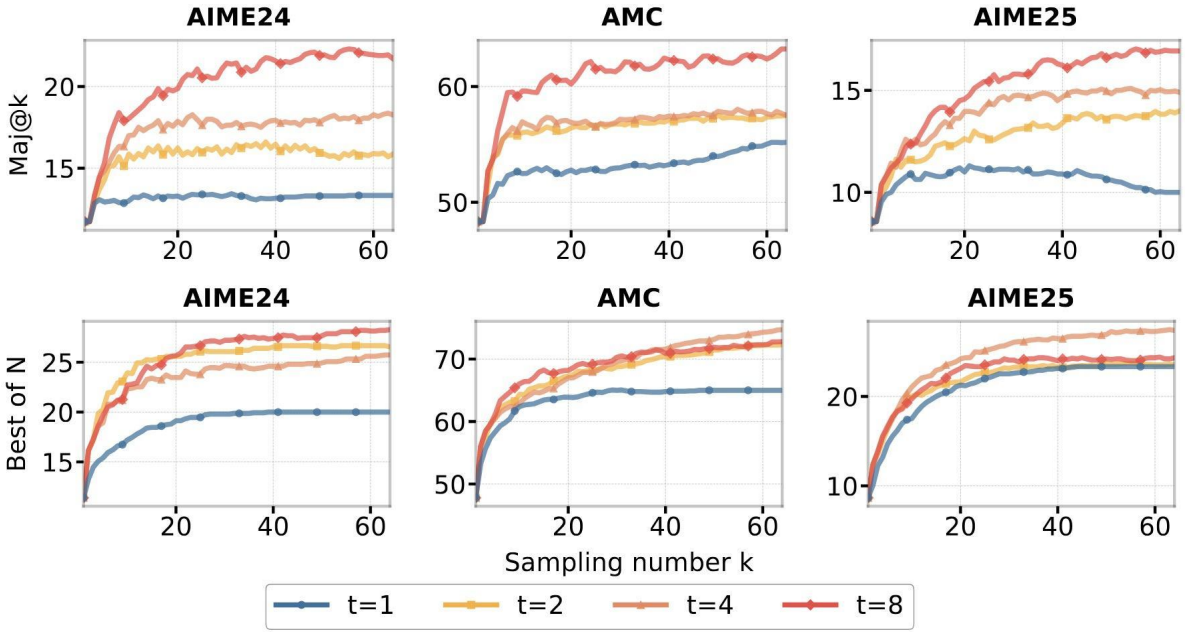


- Pass@k for different t (numbers of CKPTs) on the AIME2024, AMC, and AIME2025 benchmarks when using Temporal Sampling.
- The case t=1 represents the baseline of standard Pass@k sampling on the final CKPT.
- Temporal Sampling with t=8 outperforms the baseline by more than 19, 13, and 4 percentage points on AIME2024, AMC, and AIME2025, respectively, when totally sampling 64 responses.

Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Temporal Sampling

Temporal Sampling has better Majority Voting and Best-of-N performance than sampling only on the final checkpoint.



- t=1 represents standard Majority Voting/ Best-of-N on the final CKPT.
- Temporal Sampling outperforms the baseline by up to 8 points given the same # of sampled responses.

Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Open Discussion of RL ’s Role: Amplify vs Discovery

Debate of RL's Role: Amplify vs Discovery

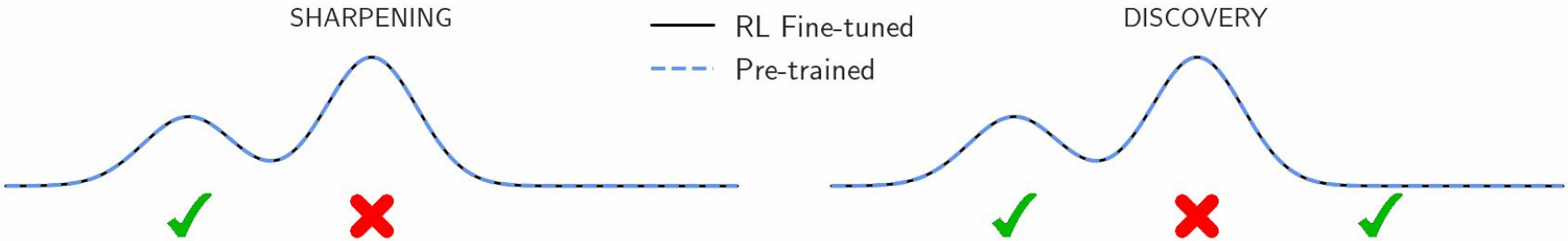
- RL amplifies reasoning patterns that was already seen during pre-training.
- It only sharpens what it may already know without crossing reasoning boundaries.
- Improvements come from selective emphasis, not the creation of entirely new ideas.

🔍 Amplify

- RL can drive the model beyond its learned priors into unexplored reasoning territory.
- By valuing successful novelty, it encourages unexpected combinations of strategies.
- This exploratory pressure yields new solutions that the base model rarely produced.

🔬 Discovery

Amplify vs Discovery



If we view it from the distribution perspective

🔍 Amplify

🔬 Discovery

https://pinnate-flare-8f3.notion.site/Sharpening-or-Discovery-RL-or-Meta-RL-How-RL-Improves-LLM-Reasoning-20628c119540805cac48e8492638d88e

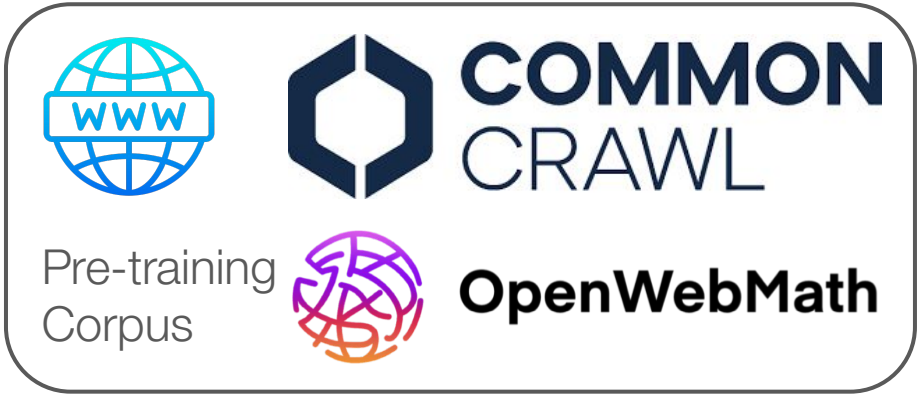
Li, Yuetai*, …, **Xiang Yue**, and Radha Poovendran. "Temporal Sampling for Forgotten Reasoning in LLMs." *arXiv 2025* ***(*: my advisee)***

Amplify

RL Strengthens Reasoning Patterns in Pre-training

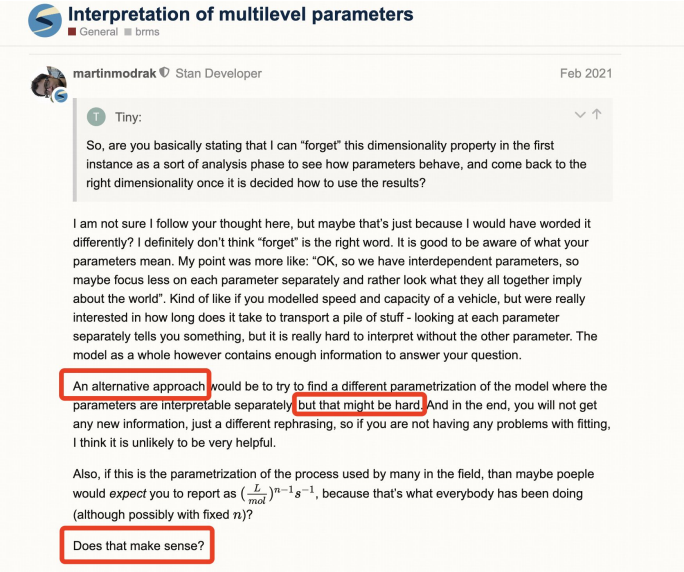
"Aha" Phrases

- "Let's think step by step."
- **"Alternatively, ..."**
- "Breaking it down step by step..."
- "Thinking about it logically, first..."
- "Step 1: Let's figure out the starting point."
- "If we follow the steps carefully, we get..."
- "To solve this, lets analyze it piece by piece."
- "Going through this systematically, we have..."
- "Okay, let's solve this gradually."
- **"Does that make sense?"**
- **"Is this correct?"**
- "Wait, does that check out?"
- "Wait, actually..."
- "Oh, hold on..."
- **"Wait a second..."**
- **"Actually, let me rethink that."**
- "Hmm, let me go back for a moment."



Search over the web pre-training corpus

Yeo*, E., Tong*, Y., Niu, M., Neubig, G., & **Yue, X.** (2025). Demystifying Long Chain-of-Thought Reasoning in LLMs. *arXiv 2025*. ***(*: my advisee)***



Interpretation of multilevel parameters

General · brms

martinmodrak Stan Developer — Feb 2021

> Tiny:
>
> So, are you basically stating that I can "forget" this dimensionality property in the first instance as a sort of analysis phase to see how parameters behave, and come back to the right dimensionality once it is decided how to use the results?

I am not sure I follow your thought here, but maybe that's just because I would have worded it differently? I definitely don't think "forget" is the right word. It is good to be aware of what your parameters mean. My point was more like: "OK, so we have interdependent parameters, so maybe focus less on each parameter separately and rather look what they all together imply about the world". Kind of like if you modelled speed and capacity of a vehicle, but were really interested in how long does it take to transport a pile of stuff - looking at each parameter separately tells you something, but it is really hard to interpret without the other parameter. The model as a whole however contains enough information to answer your question.

An alternative approach would be to try to find a different parametrization of the model where the parameters are interpretable separately but that might be hard. And in the end, you will not get any new information, just a different rephrasing, so if you are not having any problems with fitting, I think it is unlikely to be very helpful.

Also, if this is the parametrization of the process used by many in the field, than maybe poeple would *expect* you to report as $(\frac{L}{mol})^{n-1}s^{-1}$, because that's what everybody has been doing (although possibly with fixed n)?

Does that make sense?

https://discourse.mc-stan.org/t/interpretation-of-multilevel-parameters/20846

So the question is then to find the right prediction task, looking at your setup, those may include:

…

... I am not sure I follow your thought here, but **maybe that's just because I would have worded it differently**?

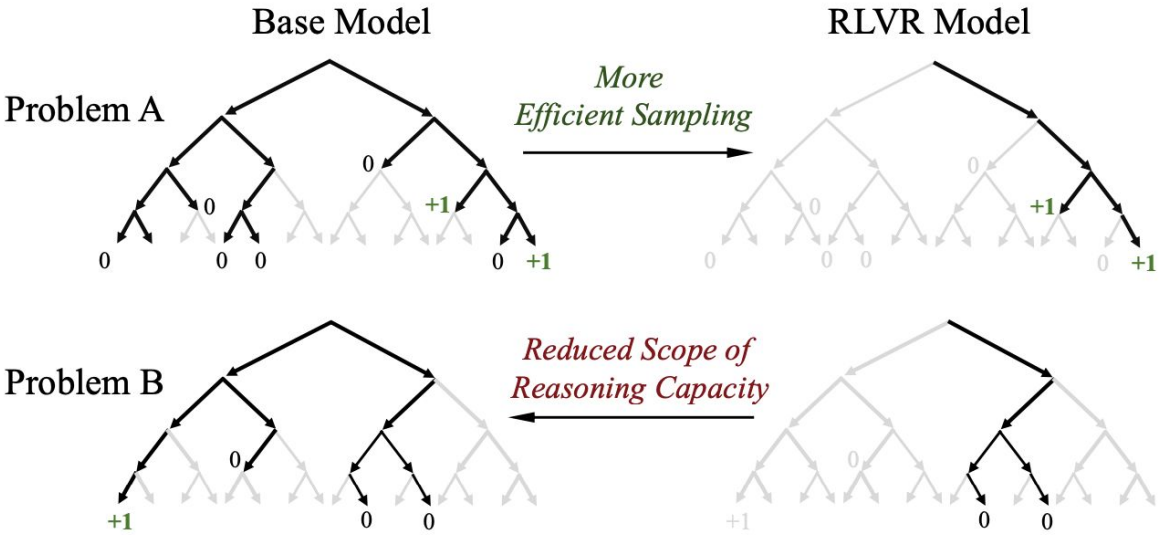
... **An alternative approach** would be to try to find a different parametrization of the model where the parameters are interpretable separately, **but that might be hard**.

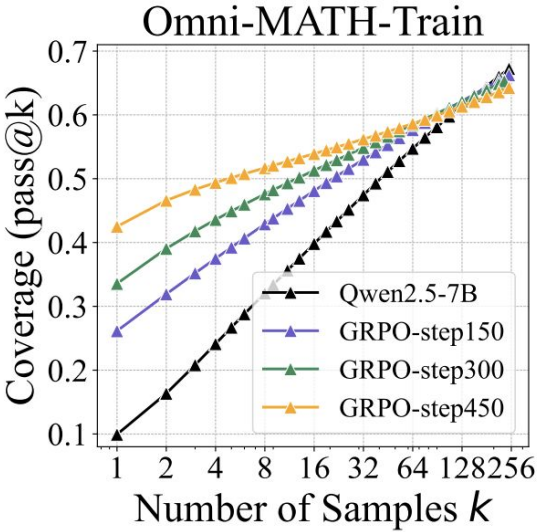
Does that make sense?

The base model might already acquire such skills during **pre-training. RL reinforces and increases the frequency of these patterns.**

Yeo*, E., Tong*, Y., Niu, M., Neubig, G., & **Yue, X.** (2025). Demystifying Long Chain-of-Thought Reasoning in LLMs. *arXiv 2025*. ***(*: my advisee)***

RL does not improve Pass@K



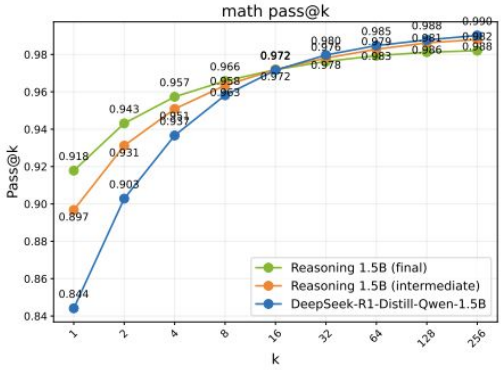


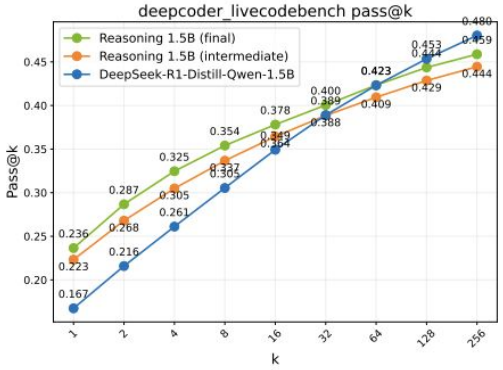
Yue, Yang, et al. "Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?." *arXiv 2025*

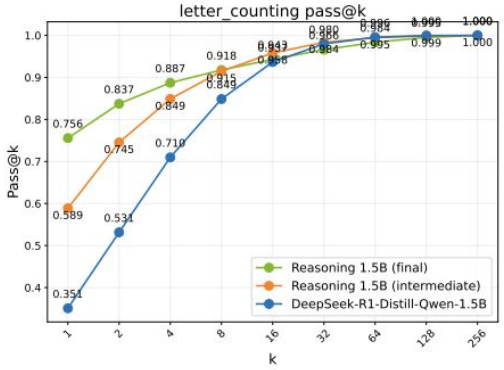
Discovery

RL does improve Pass@K (on novel complex tasks)

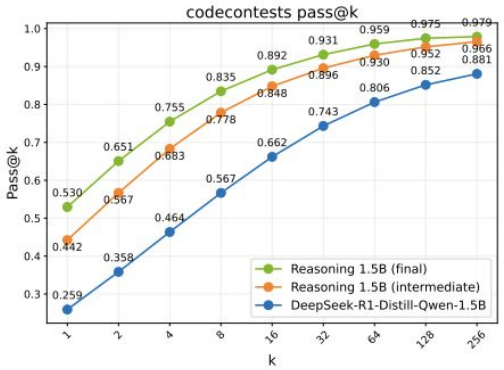
Diminish

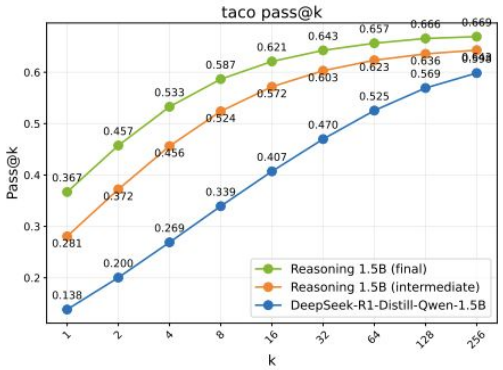


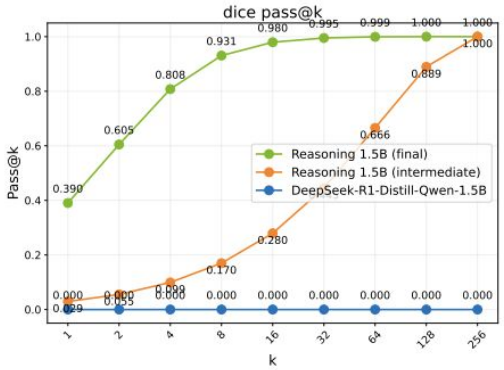




Sustained

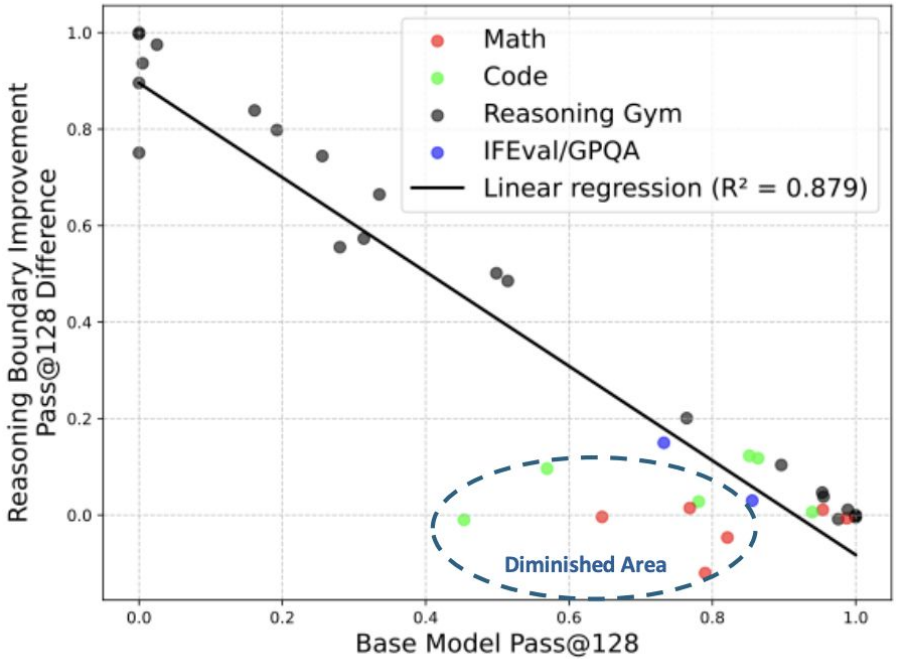


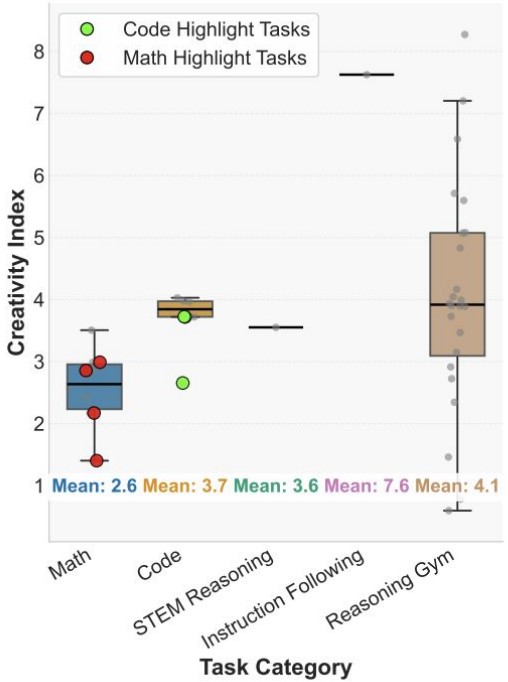




Liu, Mingjie, et al. "Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models." arXiv 2025

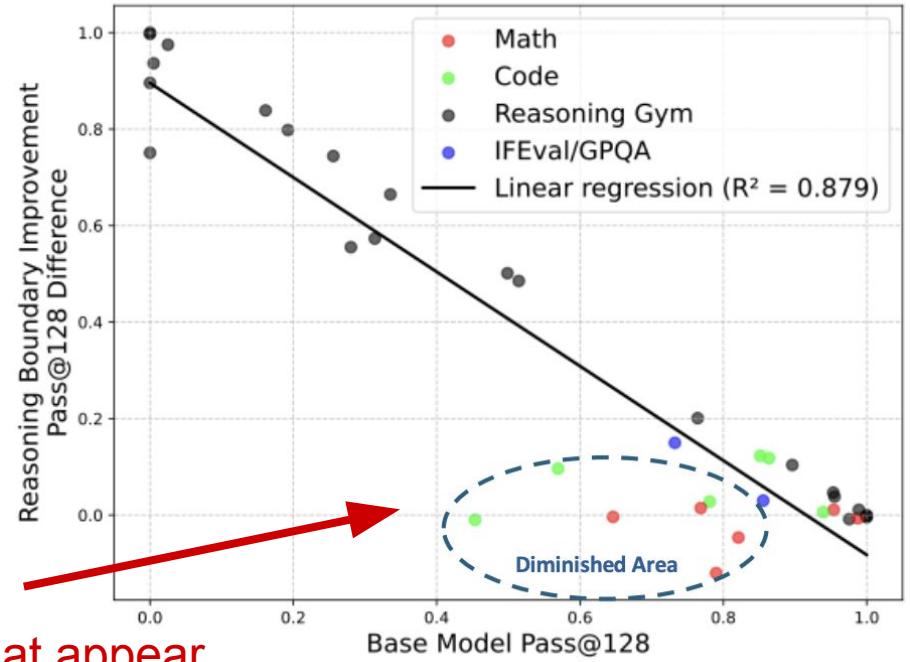
RL does improve Pass@K (on novel complex tasks)

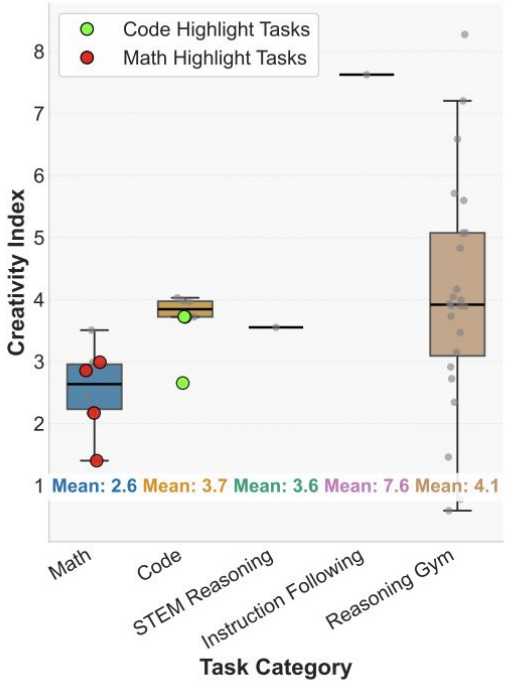




Liu, Mingjie, et al. "Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models." arXiv 2025

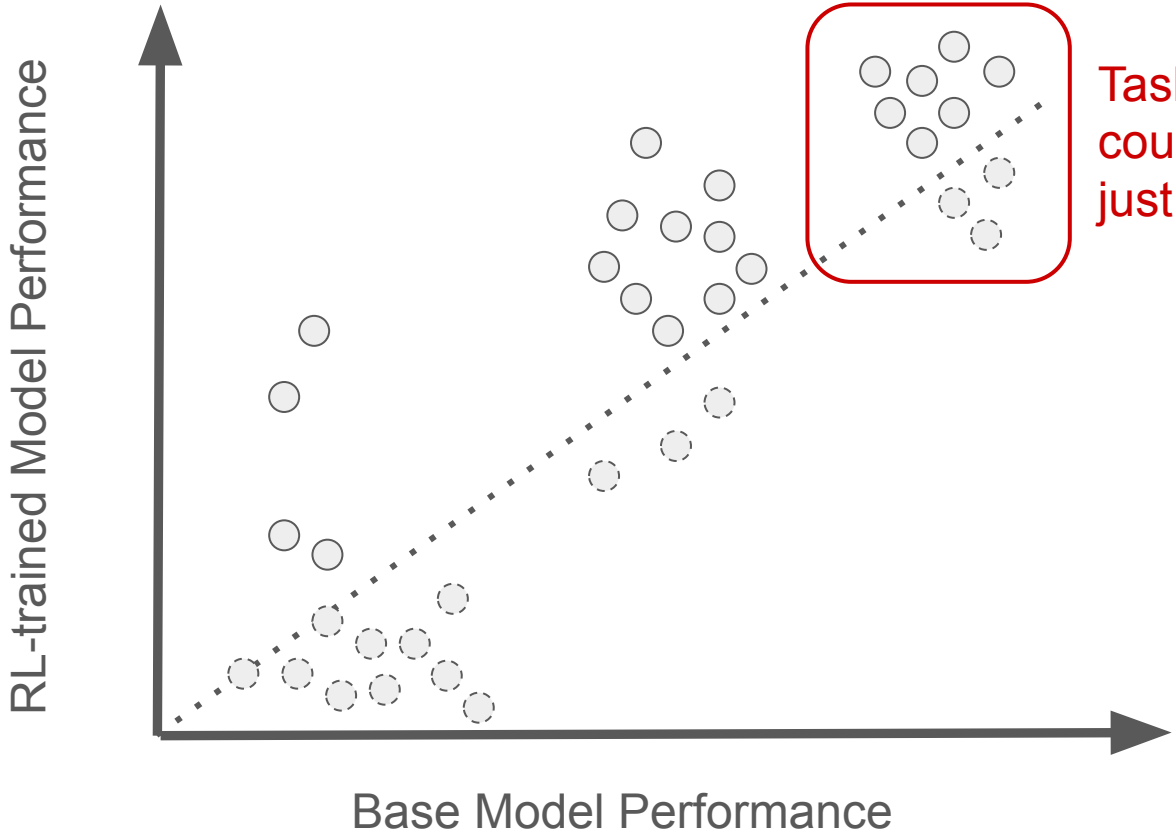
RL does improve Pass@K (on novel complex tasks)



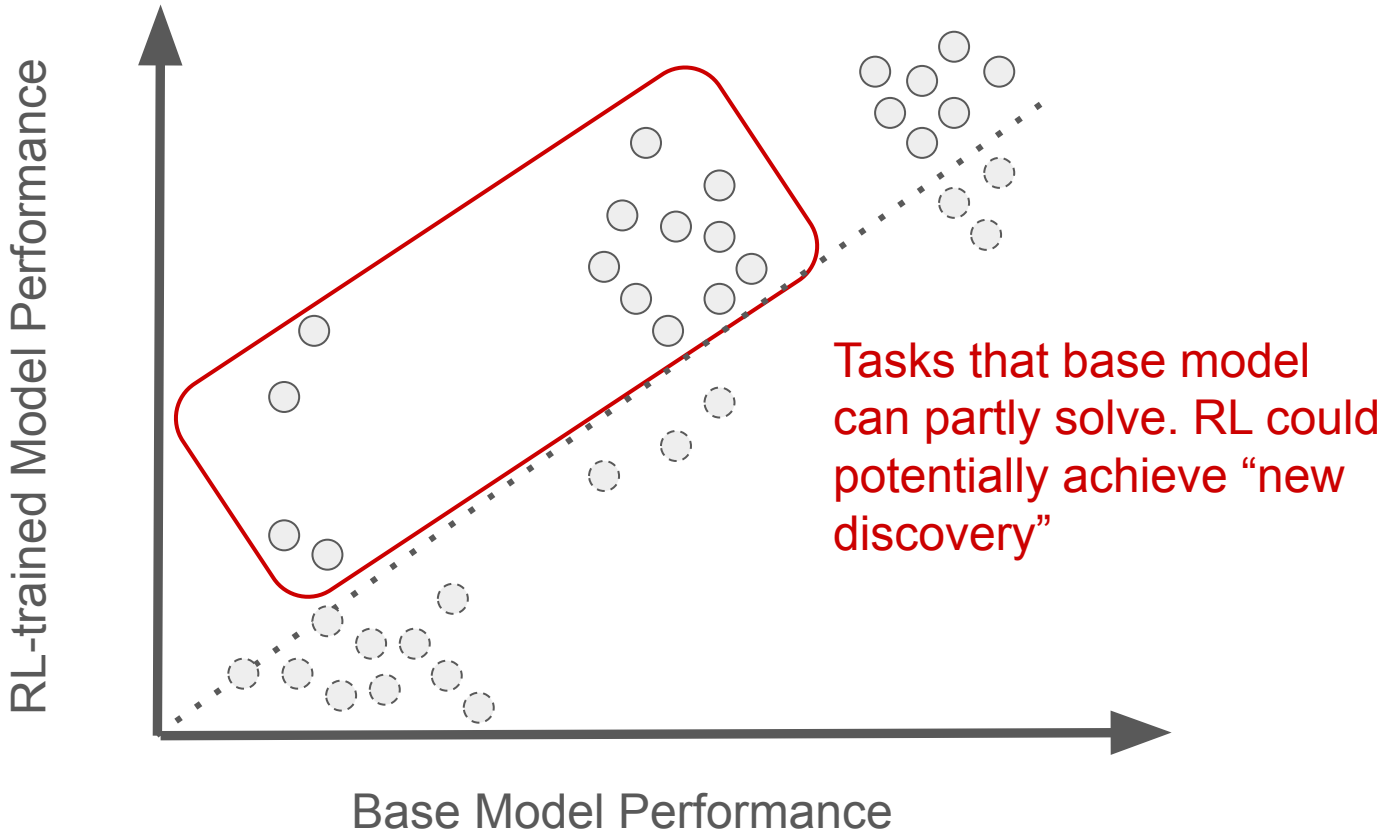


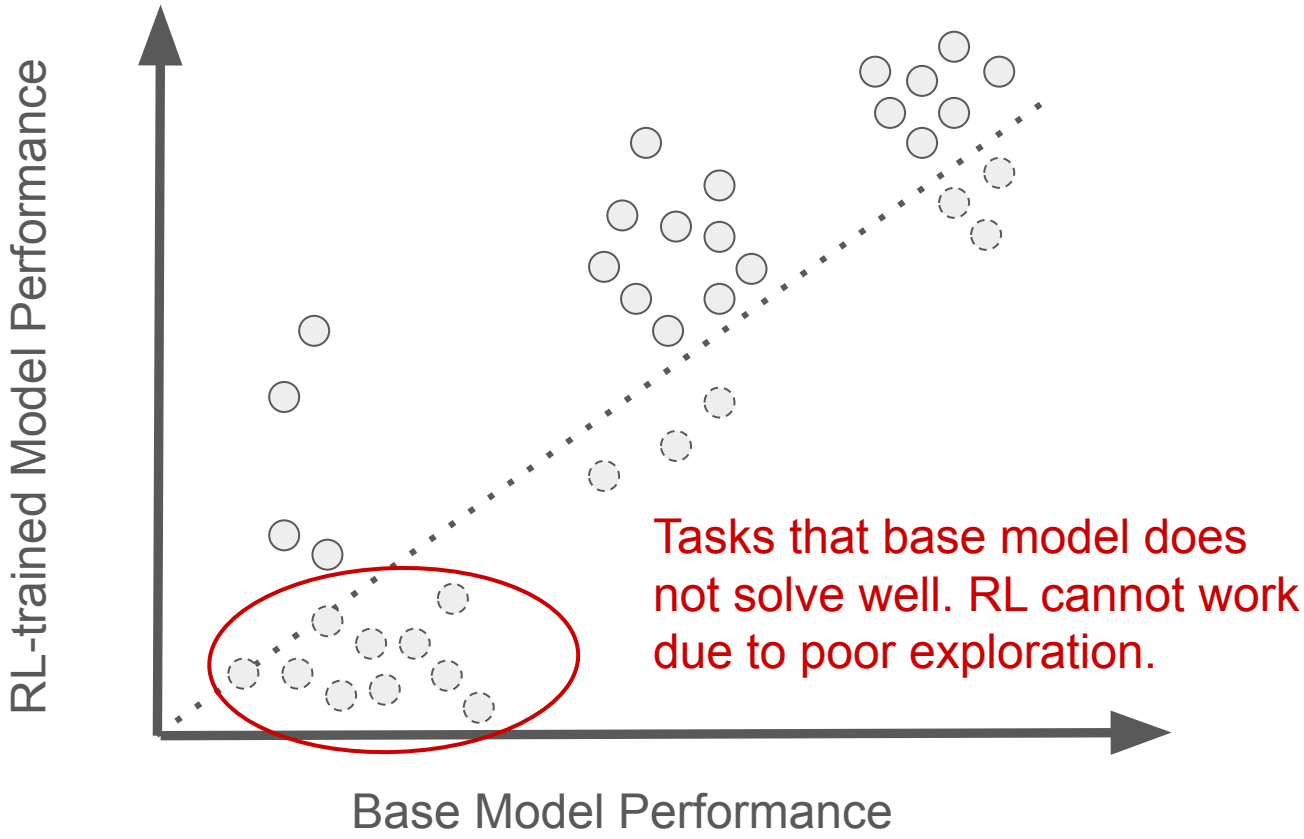
Tasks that appear frequently in pre-training

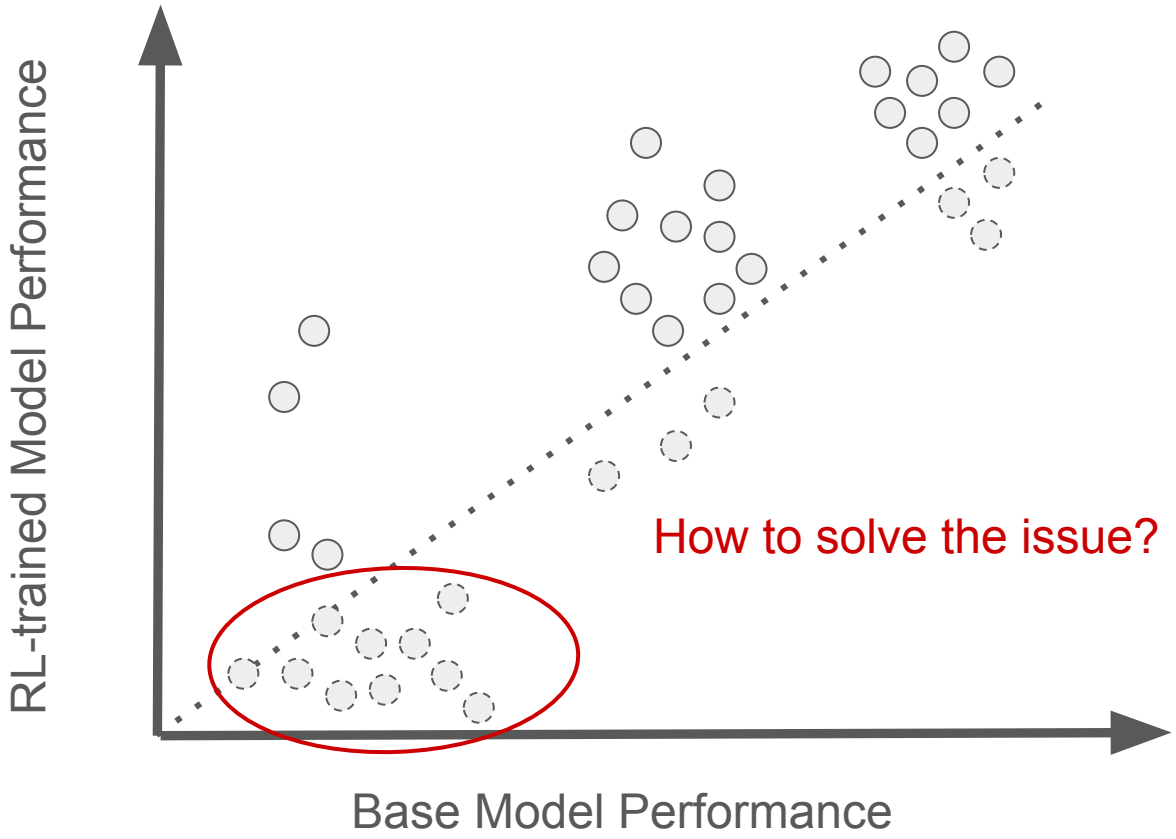
Liu, Mingjie, et al. "Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models." arXiv 2025

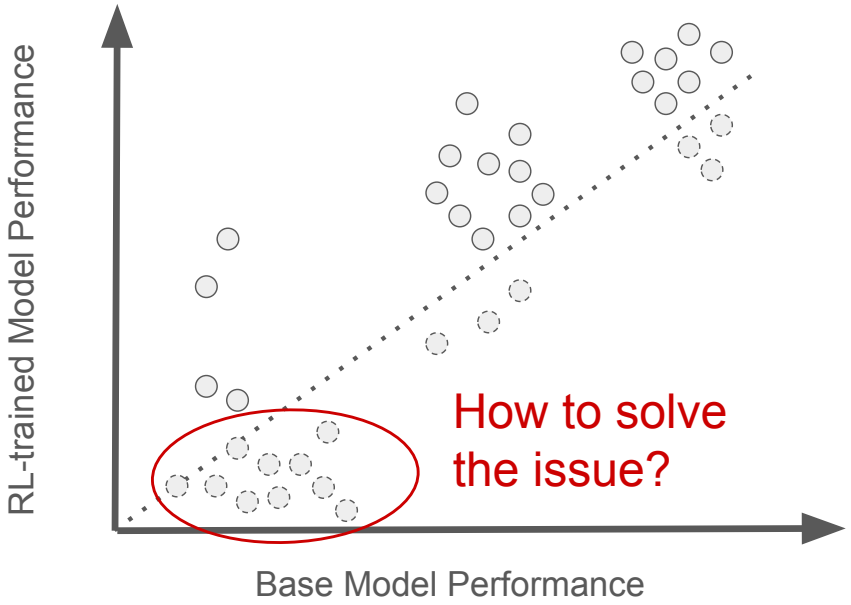


Tasks that base model could solve well. RL just elicits reasoning


RL-trained Model Performance
Base Model Performance
Tasks that base model can partly solve. RL could potentially achieve “new discovery”

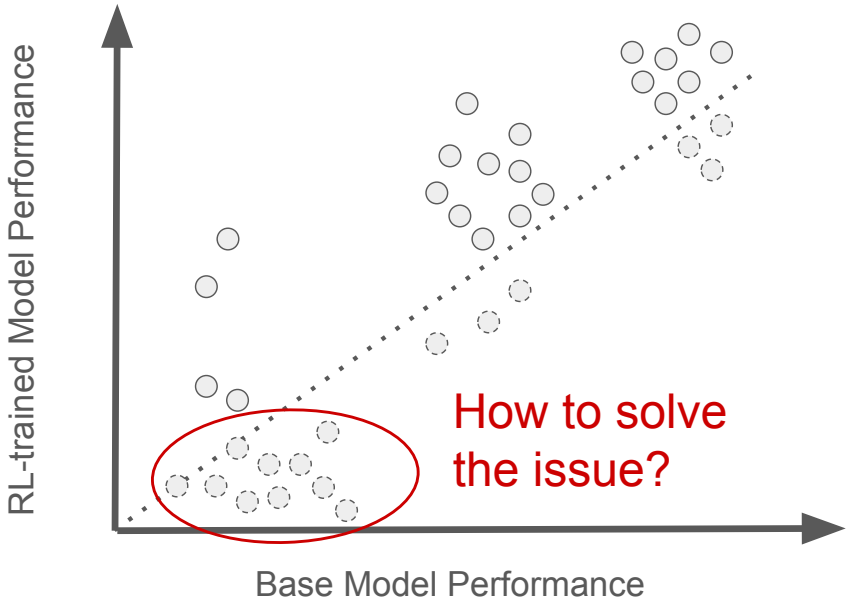

RL-trained Model Performance
Base Model Performance
Tasks that base model does not solve well. RL cannot work due to poor exploration.


RL-trained Model Performance
How to solve the issue?
Base Model Performance



Two Reasons lead to poor exploration:

- Lack of knowledge for tasks
- Lack of reasoning skills for tasks



Two Reasons lead to poor exploration:

- Lack of knowledge for tasks
- Lack of reasoning skills for tasks

Solutions:

- Pre-training / Mid-training (e.g., synthetic data for long-tail tasks).
- Off-policy RL or Hybrid RL

Is Pass@K a good metric?

What it actually measures

Coverage of correct outputs under a specific decoding policy given *k* tries.

Is Pass@K a good metric?

What it actually measures

Coverage of correct outputs under a specific decoding policy given *k* tries.

What if the model *knows but can't say?*

- A more advanced decoding strategy could solve the issue
- The model might know whether an answer is correct (e.g., good critic) but cannot synthesize from scratch

Is Pass@K a good metric?

What it actually measures

Coverage of correct outputs under a specific decoding policy given *k* tries.

What if the model *knows but can't say?*

- A more advanced decoding strategy could solve the issue
- The model might know whether an answer is correct (e.g., good critic) but cannot synthesize from scratch

Potential Solutions:

- Probing
- Test the critic ability instead of generation
- Entropy tests (correct chains vs wrong chains)

What is Intelligence?

Intelligence as a *collection of task-specific skills*

intelligence if done by humans."
--Marvin Minsky

Intelligence as a *general learning ability*

machines do tasks they have never seen"
--John McCarthy

We are mostly training AI to be strong "problem-solvers"
We might think training AI to be "general learners"

Intelligence measures a model's ability to *efficiently acquire and apply skills to achieve goals* in *novel and dynamic* environments

(My view on "Intelligence")

Acknowledgement

Yuetai Li

Maggie Huan

Tuney Zheng

Xiaoyu Xu

Yuxuan Tong

Edward Yeo

Thank you!

Email: xyue2@andrew.cmu.edu

Homepage: https://xiangyue9607.github.io/

Twitter/X: @xiangyue96