香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Improving Performance and Efficiency Through Reinforcement Learning

Weihao Zeng
HKUST CSE
April 19, 2025

Outline

1. Introduction

2. How to define good synthetic data in self-improving?

3. SimpleRL-Zoo: RL for open base models in the wild

4. Future Directions

Complex reasoning task

Complex reasoning like math and coding is an important topic in recent research

Problem

Every morning Aya goes for a 9-kilometer-long walk and stops at a coffee shop afterwards. When she walks at a constant speed of s kilometers per hour, the walk takes her 4 hours, including t minutes spent in the coffee shop. When she walks $s+2$ kilometers per hour, the walk takes her 2 hours and 24 minutes, including t minutes spent in the coffee shop. Suppose Aya walks at $s+\frac{1}{2}$ kilometers per hour. Find the number of minutes the walk takes her, including the t minutes spent in the coffee shop.

Solution 1

$\frac{9}{s}+t=4$ in hours and $\frac{9}{s+2}+t=2.4$ in hours.

Subtracting the second equation from the first, we get,

$$\frac{9}{s}-\frac{9}{s+2}=1.6$$

Multiplying by $(s)(s+2)$, we get

$$9s+18-9s=18=1.6s^2+3.2s$$

Multiplying by 5/2 on both sides, we get

$$0=4s^2+8s-45$$

Factoring gives us

$(2s-5)(2s+9)=0$, of which the solution we want is $s=2.5$.

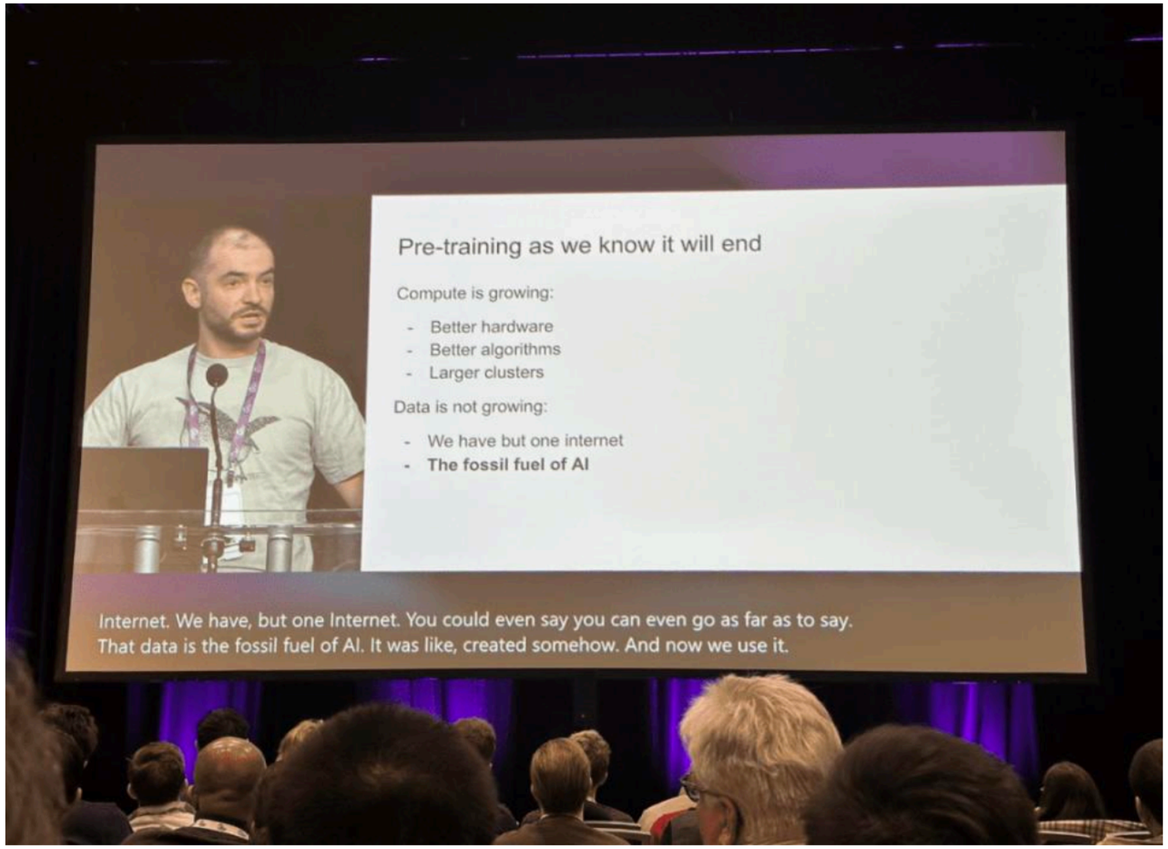
Substituting this back to the first equation, we can find that $t=0.4$ hours.

Lastly, $s+\frac{1}{2}=3$ kilometers per hour, so

$\frac{9}{3}+0.4=3.4$ hours, or $\boxed{204}$ minutes

Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.

Why RL?



Data is not growing and we are hitting the data wall

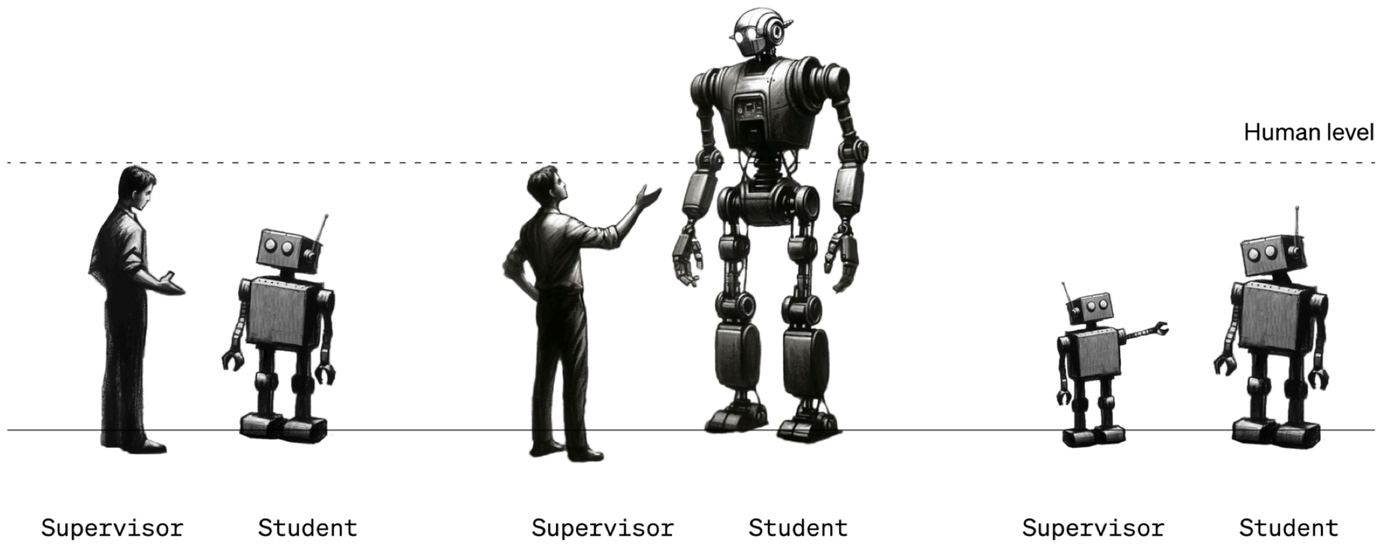
Synthetic data is becoming the future fuel to power AI

Synthetic Data in Post-Training

- Synthetic data from stronger teacher
- Synthetic data from weaker teacher
- Synthetic data from self

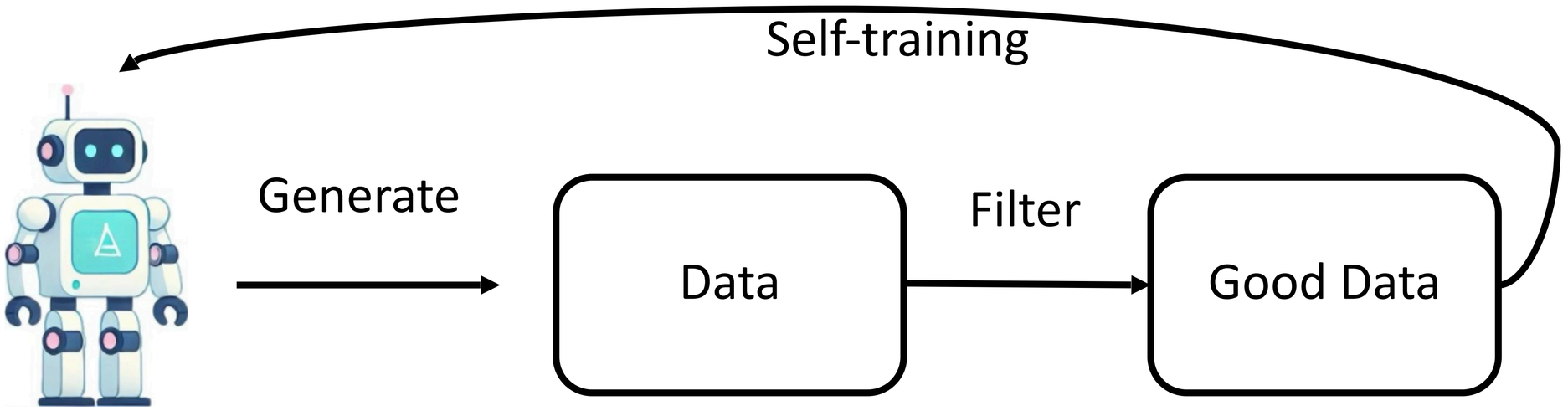
Requires existence of a stronger teacher

Requires existence of a stronger base model



[1] Burns et al. Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. 2023

Synthetic Data from Self



Each iteration faster and faster (generate less and update less)

Offline → Online

Traditional Self-Improving

(Online) Reinforcement Learning

Why RL?

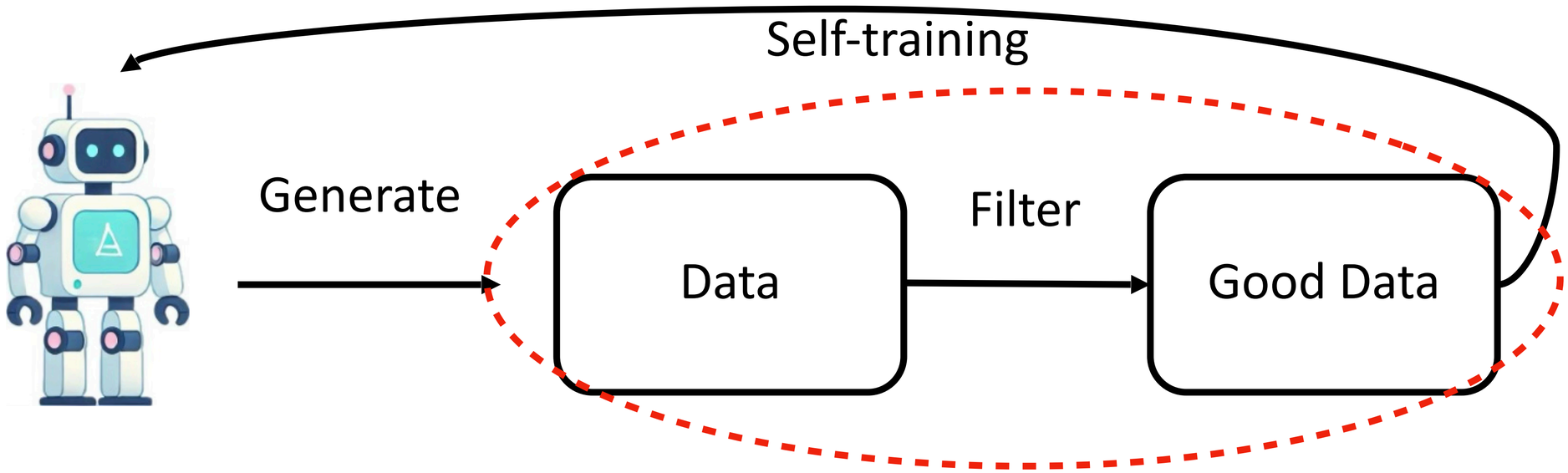
SFT: Imitation learning, requires higher quality data

Hard to surpass the teacher

RL: Encourages free exploration, with higher potential

Why RL?

RL is a form of data synthesis to overcome the human data limitation



How to manipulate the synthetic response quality?

Outline

1. Introduction

2. How to define good synthetic data in self-improving?

3. SimpleRL-Zoo: RL for open base models in the wild

4. Future Directions

B-STaR: Monitoring and Balancing Exploration and Exploitation in Self-Taught Reasoners

Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, Junxian He

ICLR 2025

Strong2Weak Distillation v.s. Self-Improving

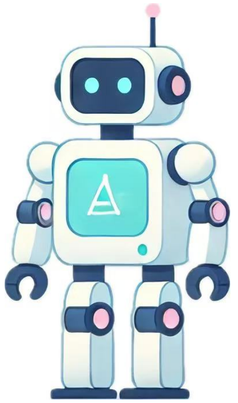
Generate

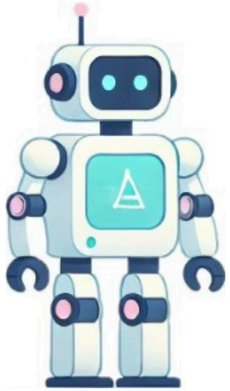
Data

Filter

Good Data

Train





Generate

Data

Filter

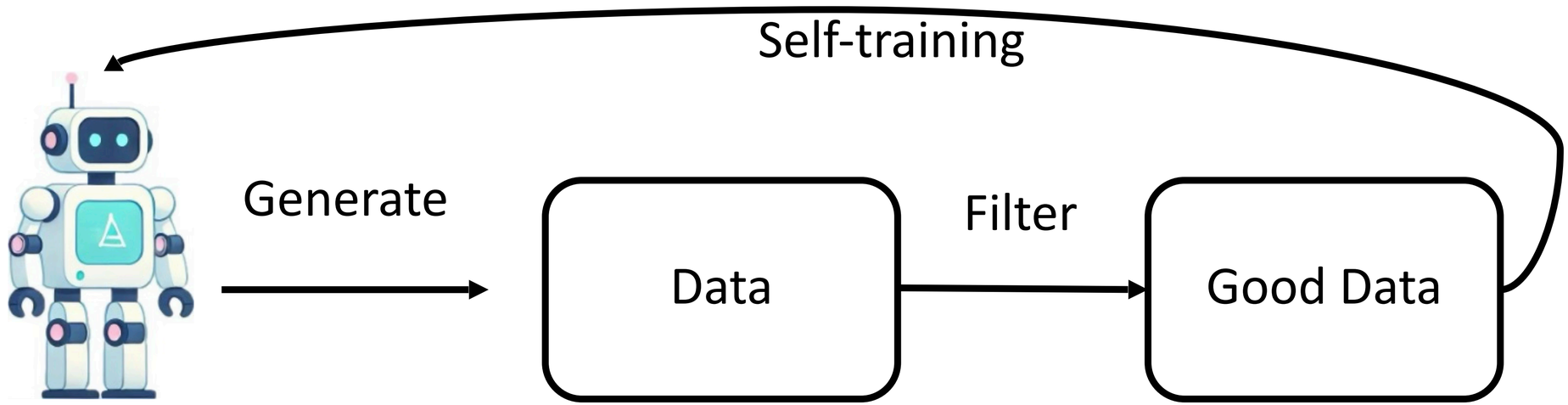
Good Data

Self-training

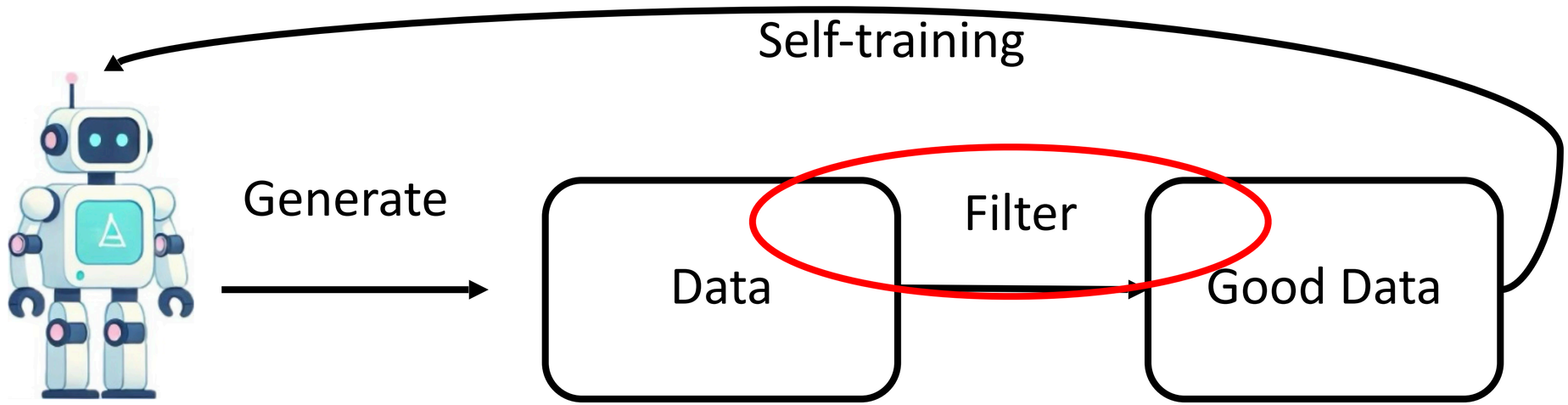
Self-Improving

Reinforcement Learning

Why can Self-Training Work?



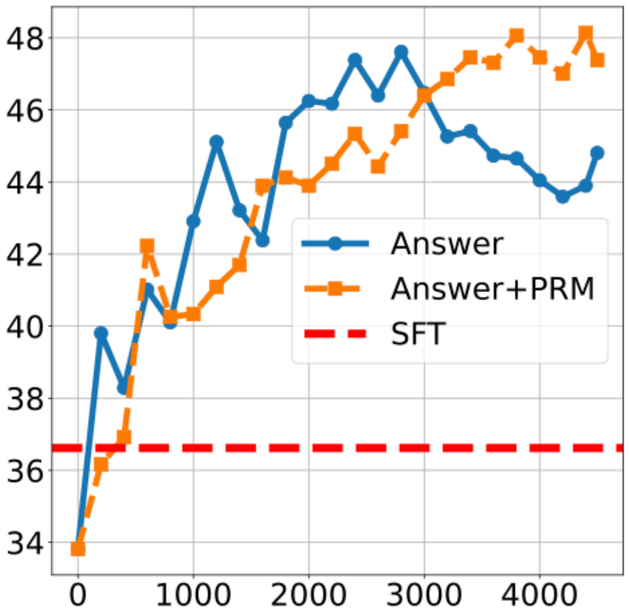
Why can Self-Training Work?



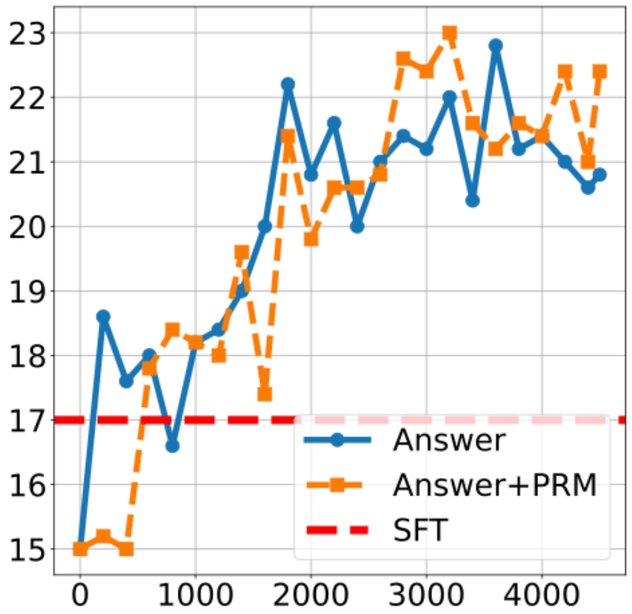
The filter step provides external signal/supervision

We experiment with 1. final answer; 2. final answer + process-based reward model (PRM)

However, the current situation of self-improving is…



(a) Pass@1 on GSM8K



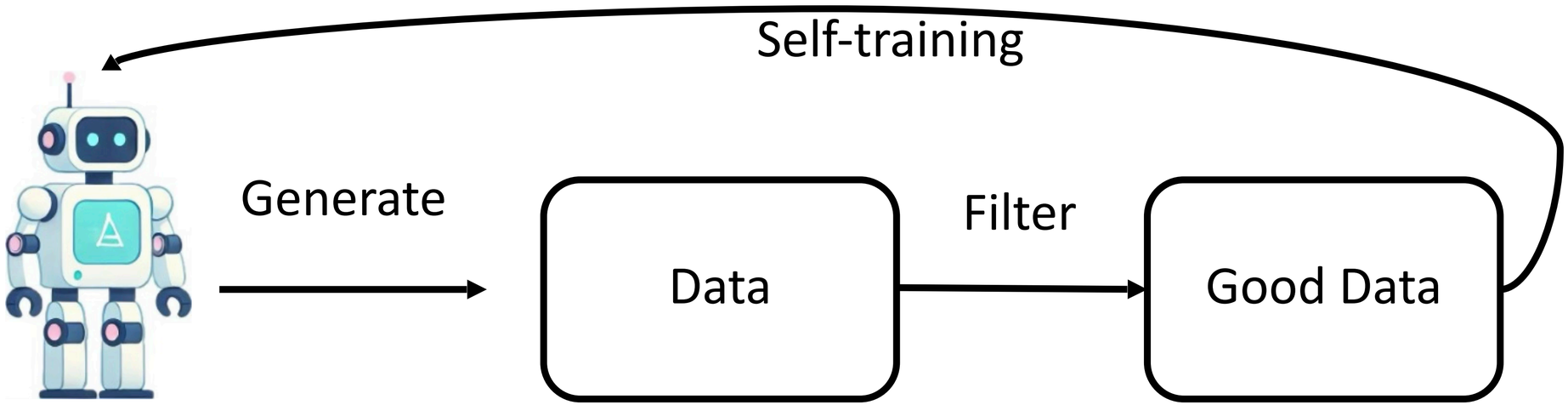
(b) Pass@1 on MATH

Performance saturates quickly

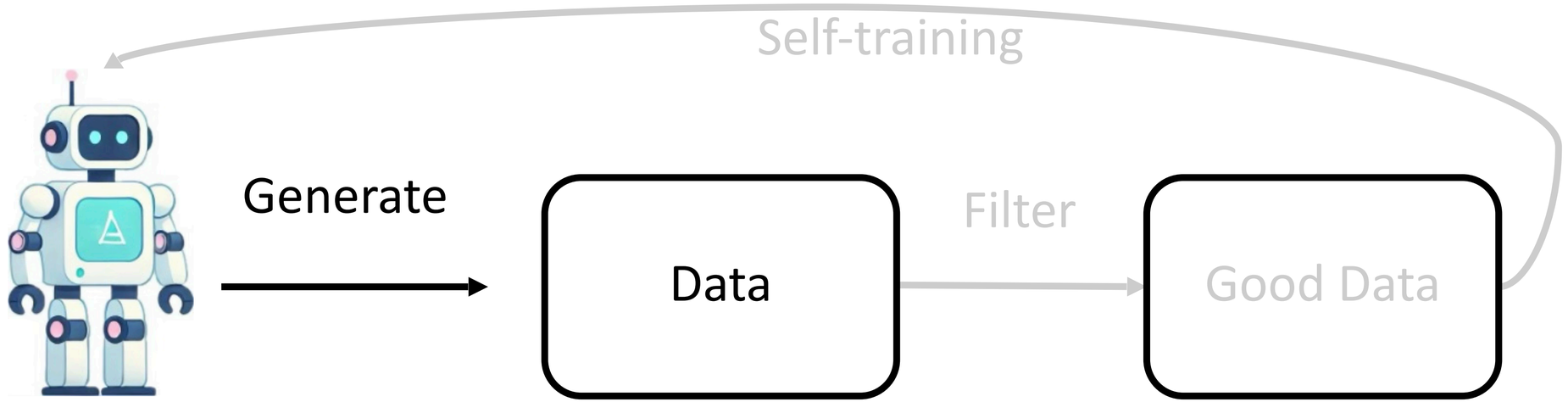
NO scaling law

Note: this is the self-improving paradigm before R1 comes out, the problem is the model generation/data synthesis is not frequent enough

What Matters in Self-Improving

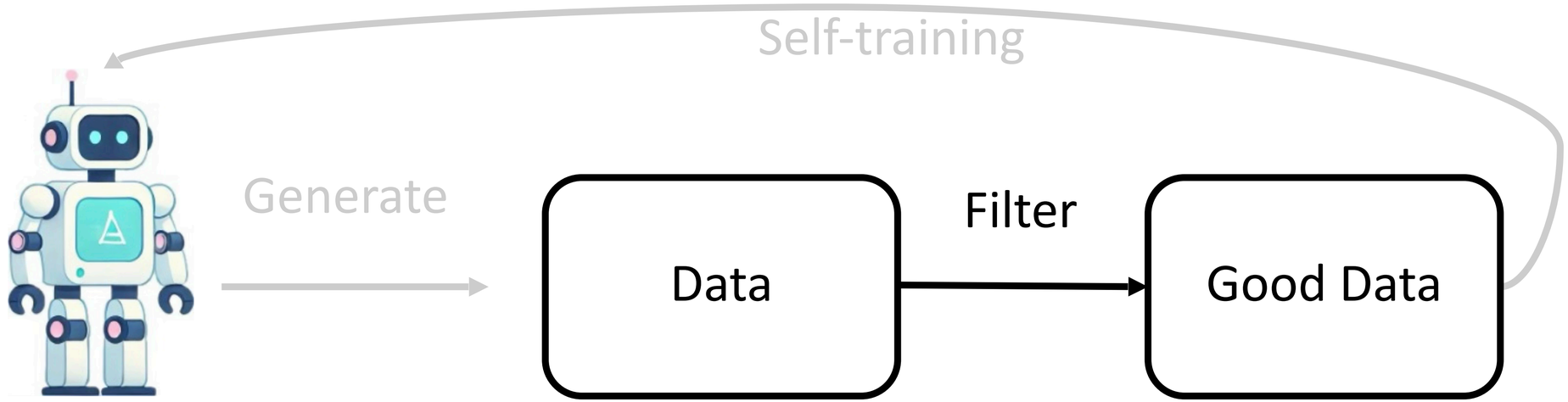


What Matters in Self-Improving



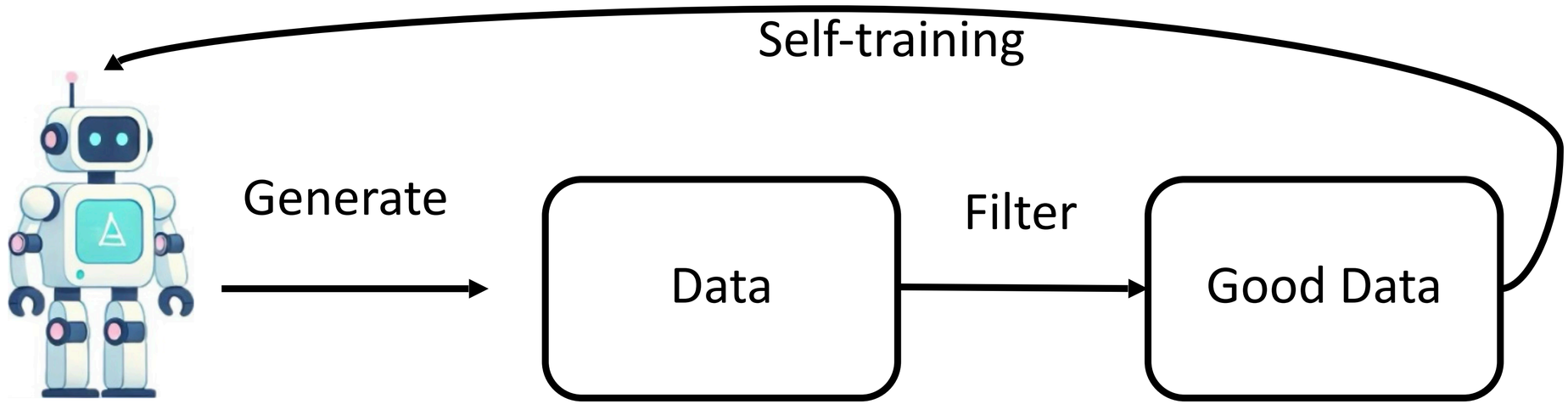
The model needs to be able to generate good responses when sampling multiple candidates

What Matters in Self-Improving



The reward function needs to be able to distinguish good responses from bad ones

What Matters in Self-Improving



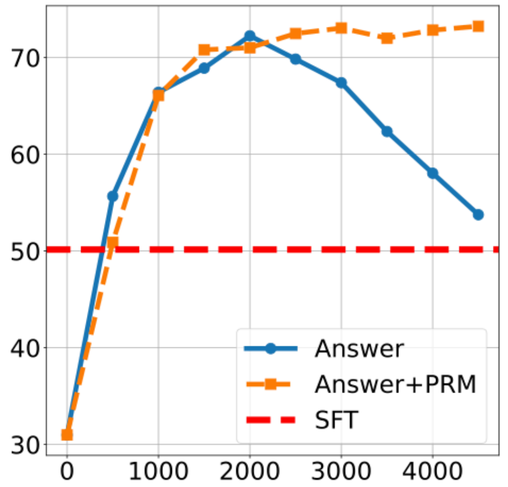
1. The model needs to be able to generate good responses when sampling multiple candidates — Exploration

2. The reward function needs to be able to distinguish good responses from bad ones — Exploitation

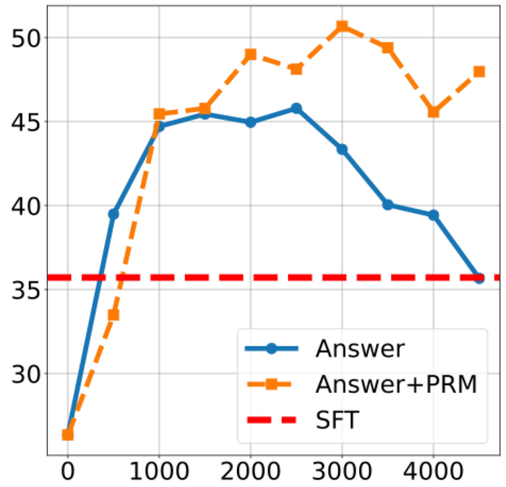
Self-improving will not work if either exploration or exploitation lags behind – balance is important!

However.. the ability of exploration and exploitation are dynamically changing, are they balanced in standard self-improving?

Exploration



(c) Diversity on GSM8K



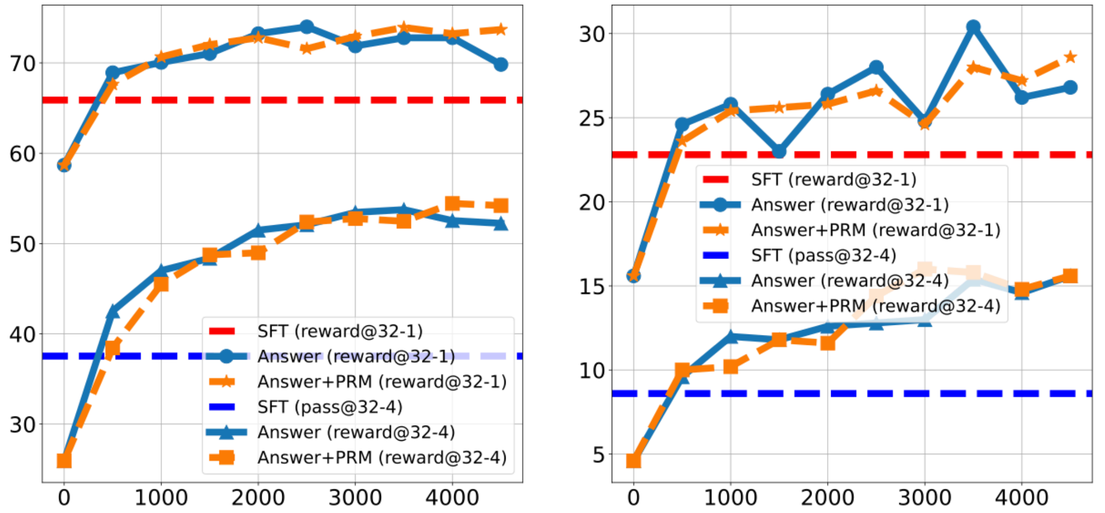
(d) Diversity on MATH

training steps

Exploration increases first, then saturates or drops

Model's exploration decides the skyline

Exploitation



(c) Reward@K-S GSM8K (d) Reward@K-S MATH

Exploitation saturates as well, but at a slower pace. It seems exploration lags behind!

How to measure the combined effect of exploration and exploitation?

Intuitively, it is like measuring the quality of the selected responses?

What Makes Good Training Data?

1. The number of correct responses among the selected data should be large enough, otherwise we don't have enough good training data

2. The ratio of correct responses among the selected data should be high, otherwise the training data is too noisy

What Makes Good Training Data?

n_i': number of correct selected responses

n_i: number of selected responses

$n^\star$: a minimum number of correct responses that we want

$$qe_i = \underline{\min\left(\frac{n_i'}{n^\star},\ 1\right)} \cdot \frac{n_i'}{n_i}$$

Discount factor to penalize insufficient correct responses

What Makes Good Training Data?

n_i': number of correct selected responses

n_i: number of selected responses

$n^\star$: a minimum number of correct responses that we want

$$qe_i = \min\left(\frac{n_i'}{n^\star},\ 1\right) \cdot \frac{n_i'}{n_i}$$

Ratio of correct responses among all the responses

What Makes Good Training Data?

n_i': number of correct selected responses

n_i: number of selected responses

$n^\star$: a minimum number of correct responses that we want

$$qe_i = \min\left(\frac{n_i'}{n^\star},\ 1\right) \cdot \frac{n_i'}{n_i}$$

It is like a direct measure of the current, rollout data contribution

How can we maximize the query effect?

What factors influence exploration and exploitation?

Exploration: sample size K, rollout temperature t

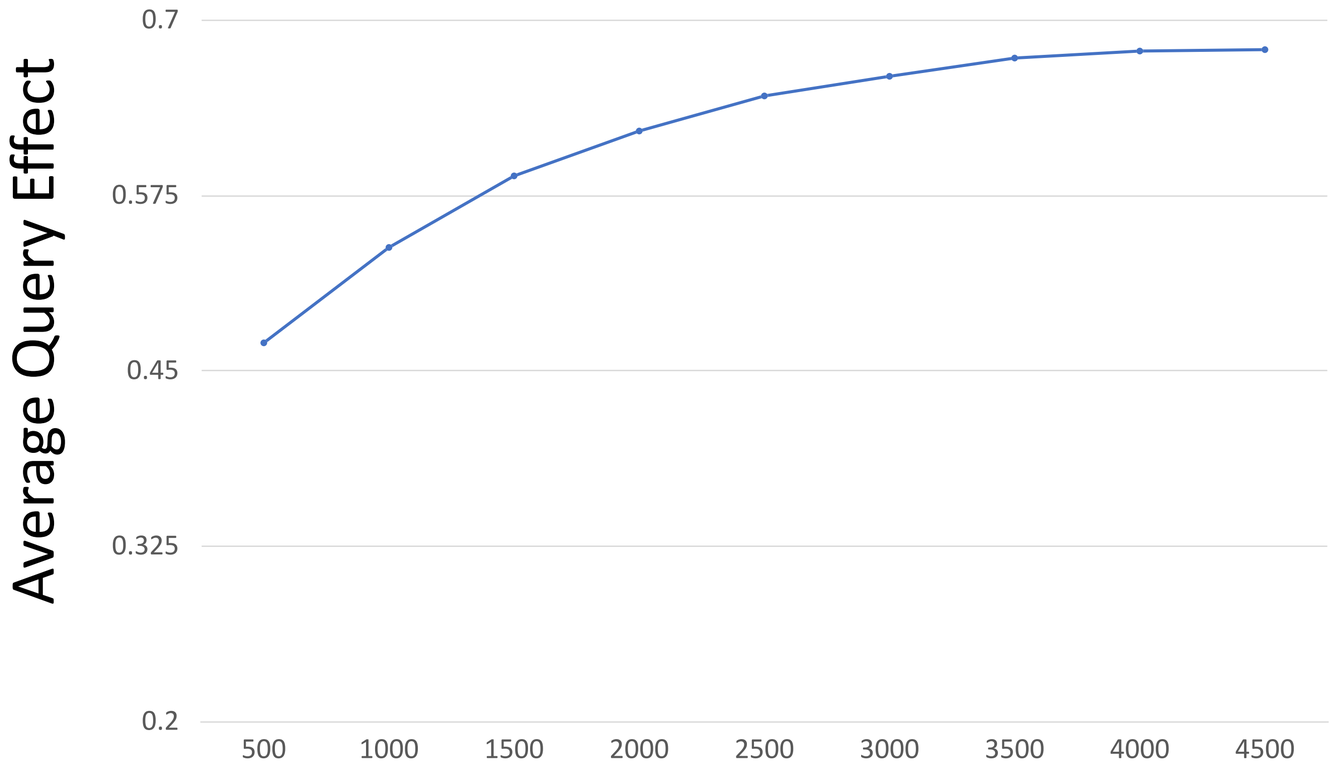
Exploitation: reward selection threshold

B-STaR (Balanced STaR): for every T training steps, we dynamically change sample size, rollout temperature, and reward threshold to maximize average query effect scores on a small training subset

We always want to maximize the rollout data contributions

This is different from common practice where these configurations are treated static

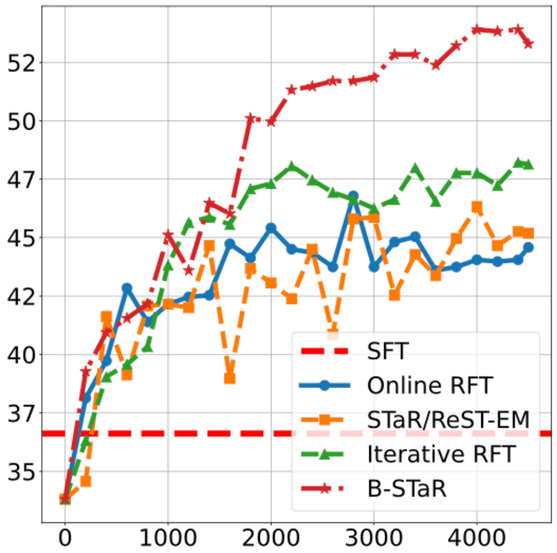
Query Effect Change of B-STaR



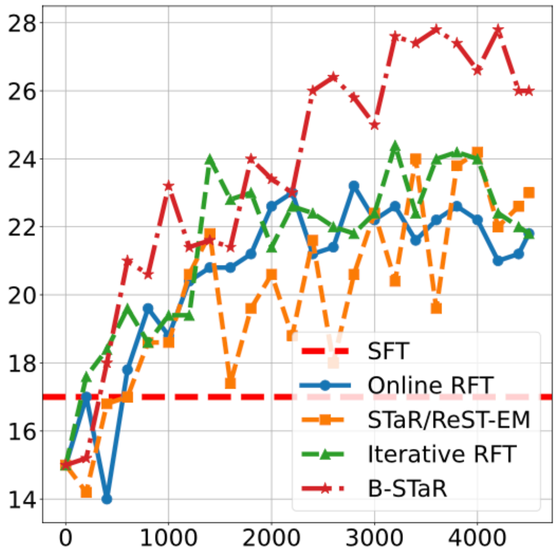
Monotonically increasing

Main Results

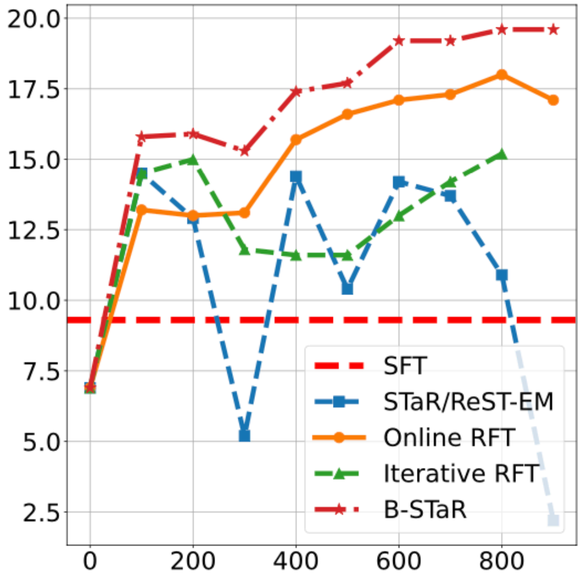
Pass@1 Accuracy



(a) GSM8K



(b) MATH



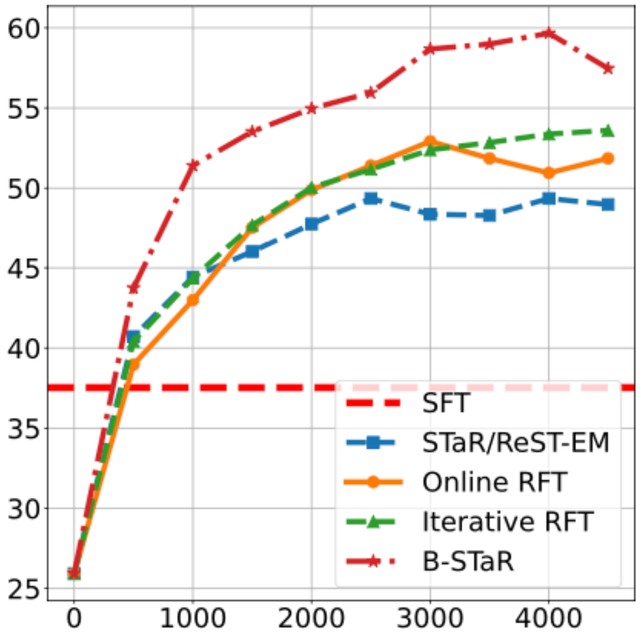
(c) APPS

training steps

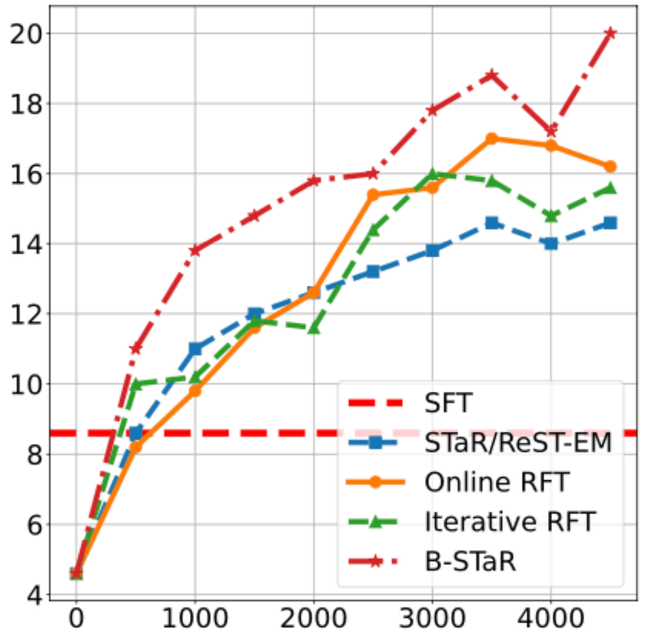
Significantly better scaling

Exploration is also Improved

Pass@32 Accuracy



(a) GSM8K



(b) MATH

How does B-STaR Change the Configurations?

Step	500	1000	1500	2000	2500	3000	3500	4000	4500
Sample Number	64	64	64	64	64	64	64	64	64
Temperature	0.5	0.8	0.9	1	1.1	1.1	0.9	1.1	1.1
Reward thresholds	0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Query Effect	0.470	0.538	0.589	0.621	0.646	0.660	0.673	0.678	0.679

Temperature increases gradually –> encourage more exploration as training runs

How does B-STaR Change the Configurations?

Step	500	1000	1500	2000	2500	3000	3500	4000	4500
Sample Number	64	64	64	64	64	64	64	64	64
Temperature	0.5	0.8	0.9	1	1.1	1.1	0.9	1.1	1.1
Reward thresholds	0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Query Effect	0.470	0.538	0.589	0.621	0.646	0.660	0.673	0.678	0.679

Sampling more is always better (we are not training more)

Outline

1. Introduction

2. How to define good synthetic data in self-improving?

3. SimpleRL-Zoo: RL for open base models in the wild

4. Future Directions

The Deep Thinking Era Arrives…..

Long Chain of Thought (CoT)

O1 Solution

Okay, let's see. So Aya goes for a 9-kilometer walk every morning and stops at a coffee shop. The problem gives two scenarios with different speeds and total times, and we need to find the total time when she walks at a different speed. Hmm.

First, I need to parse the information given. The key points are:

1. The distance is always 9 km.
2. The total time includes both walking time and t minutes at the coffee shop.
3. Two different speeds are given, each resulting in different total times, but the coffee shop time t is the same in both cases.

Wait, let me make sure. The problem states when she walks at s km/h, the total time is 4 hours (which includes t minutes). When she walks at s + 2 km/h, the total time is 2 hours and 24 minutes (which also includes t minutes). So both cases have the same t, but different walking times leading to different total times.

So, the total time is walking time plus t. Since t is in minutes, but the total time is given in hours and minutes, I need to convert everything to the same units. Maybe convert all times to minutes to avoid confusion.

.....

Since speed can't be negative, s = (5)/2 = 2.5 km/h

Wait, sqrt(49)=7, so (-2 +7)/2=5/2=2.5, yes. So s=2.5 km/h.

Now check if that makes sense.

.....

Third scenario: speed is s +0.5=3 km/h. Walking time=9/3=3 hours=180 minutes. Add t=24, total=204.

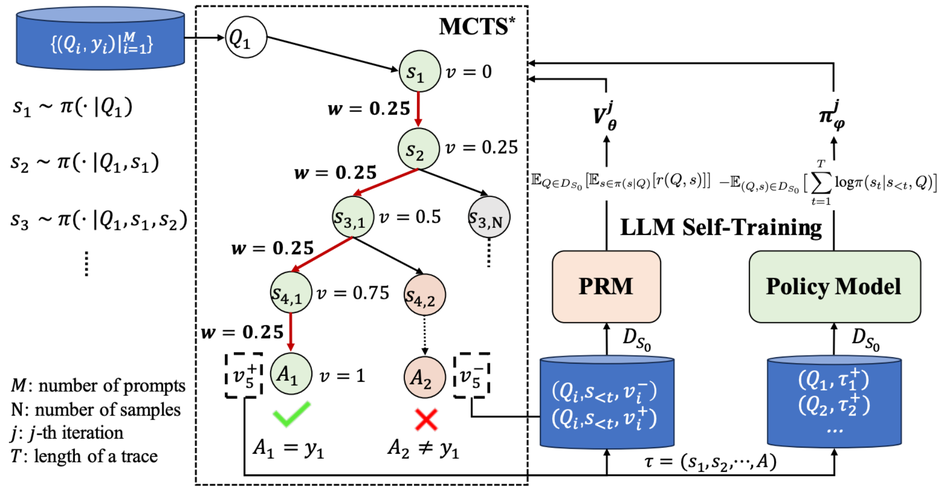
Yes, seems correct.

Long Chain of Thought (CoT)

+ certain cognitive behaviors (e.g., Self-Reflection)

Efforts for Developing O1-style Models

Many researchers are exploring possible paths towards learning o1-style models…



Tree Search

More complex during training and testing

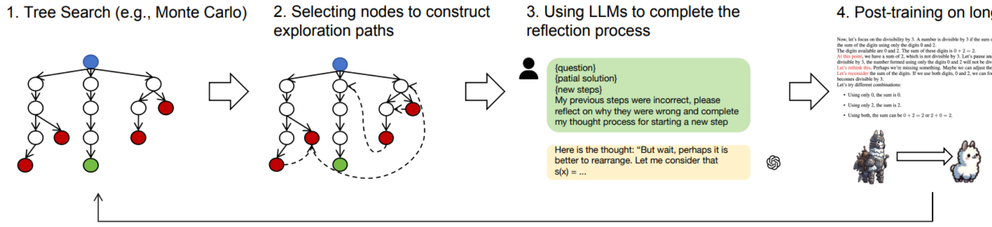


Figure 2: The framework of journey learning.

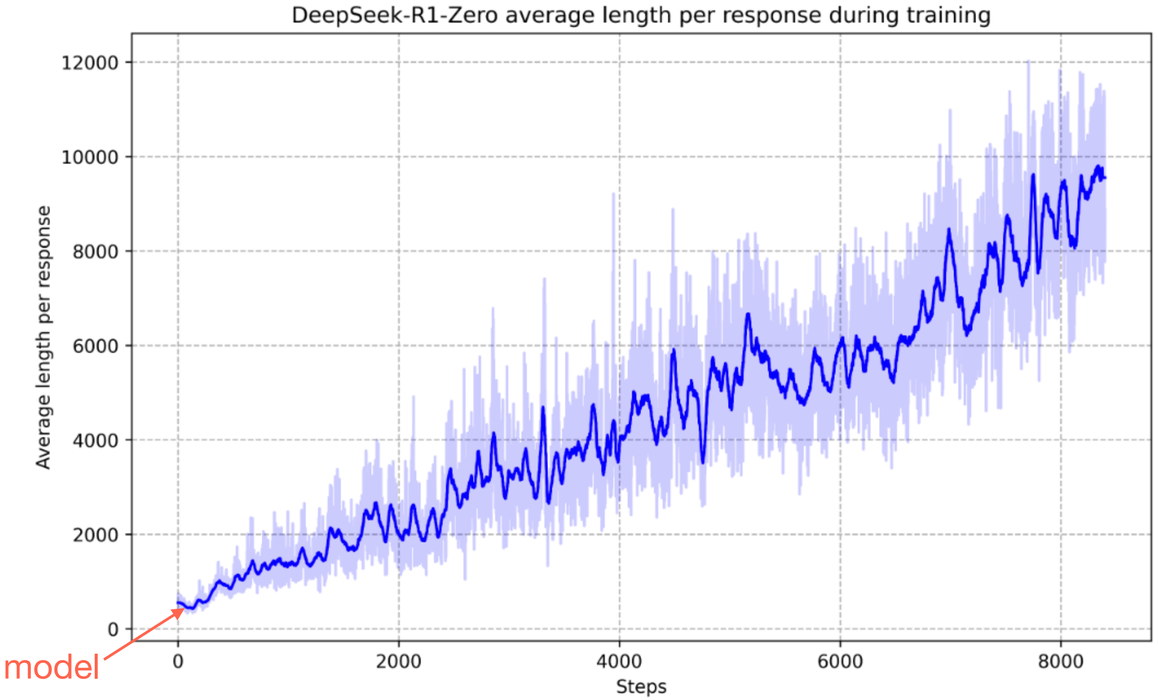
Distillation

Hard to surpass teacher model

Zhang, Dan, et al. "Rest-mcts*: Llm self-training via process reward guided tree search." *arXiv preprint arXiv:2406.03816* (2024).
Qin, Yiwei, et al. "O1 Replication Journey: A Strategic Progress Report--Part 1." *arXiv preprint arXiv:2410.18982* (2024).

Long CoT Patterns Emerging in RL Training

DeepSeek-R1 and Kimi-k1.5 choose the extremely simple recipe — Reinforcement Learning



Model naturally develops long CoT during RL

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Self-Reflection Emerging in RL Training

Example:

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 $\cdots$

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

Rearrange to isolate the inner square root term:

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

$\ldots$

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 $\cdots$

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: $\cdots$

$\ldots$

Without tree search and reward model, long CoT and self-reflection just naturally emerge.

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

RL Training is not a new thing

But the emergence of
Long CoT plus Self Reflection is new

DeepSeek R1 Two Main Parts

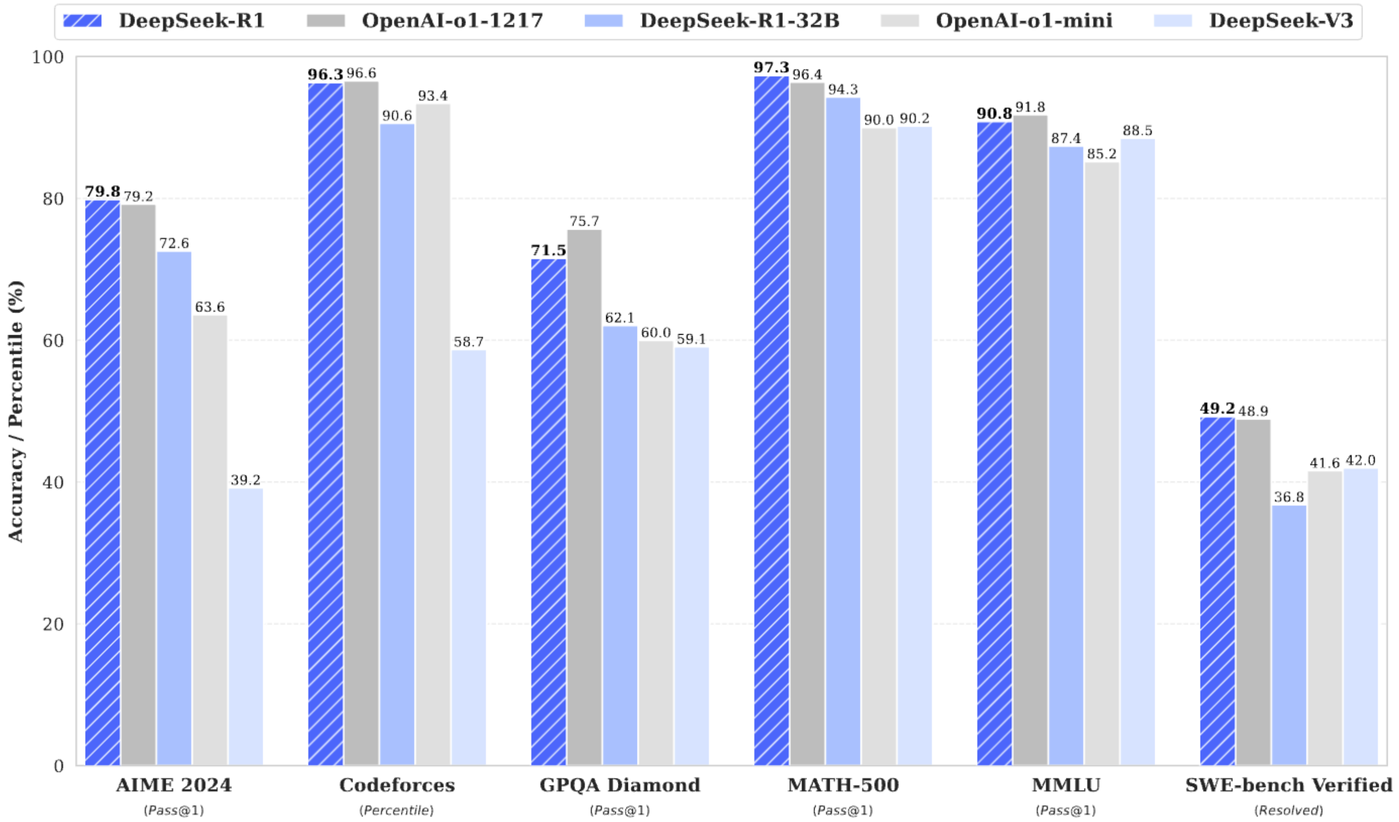
Deepseek's Paper includes two main parts:

- Deepseek R1 Zero: Just RL using the math and code data.
- Deepseek R1: SFT + RL . Some SFT data is from R1-Zero.
  - Accelerate the RL training.
  - More user-friendly by adding some non-reasoning data.

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

DeepSeek R1's Amazing Performance

Performance:



DeepSeek-R1 achieves performance comparable to OpenAI-o1 on reasoning tasks.

Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

Deepseek-R1 conducts experiments on huge models in a large-scale RL setting. What about small model and limited data?

Does RL works for small model and limited data?

Does long CoT and self-reflection emerge in this case?

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

SimpleRL: Emerging Reasoning with Reinforcement Learning is Both Effective and Efficient

Weihao Zeng*, Yuzhen Huang*, Wei Liu, Keqing He, Qian Liu, Zejun Ma, Junxian He*

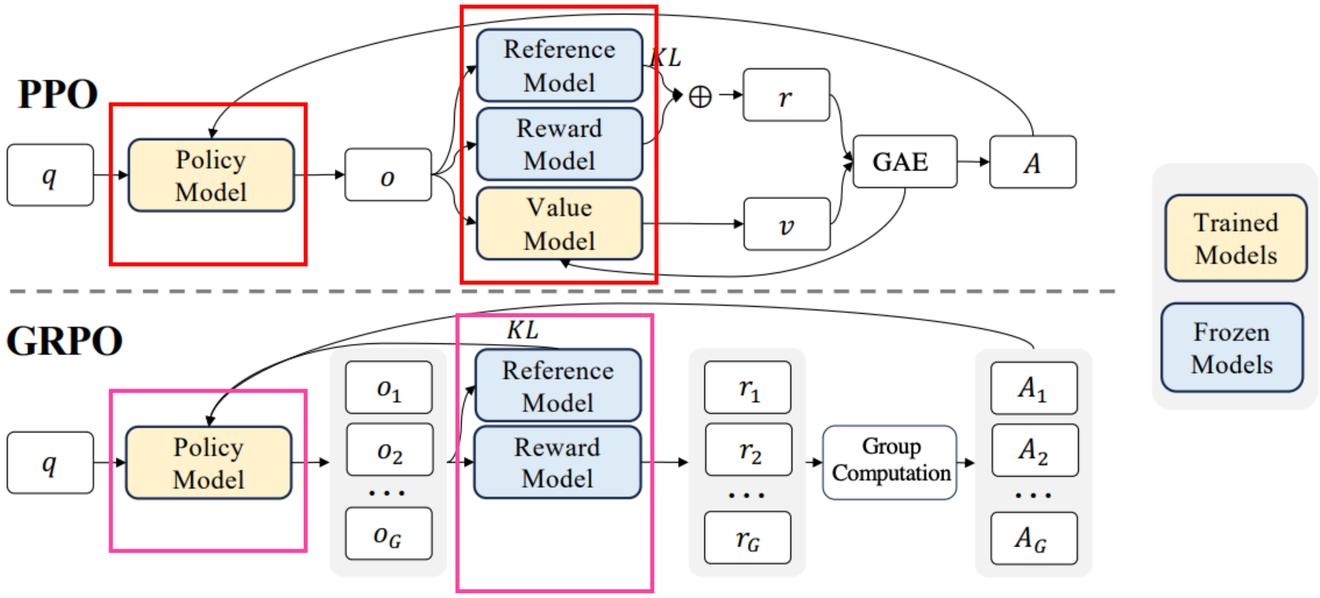
** many of the experiments were developed and performed by us independently before DeepSeek-R1's release.*

* : Project lead

The SimpleRL Recipe

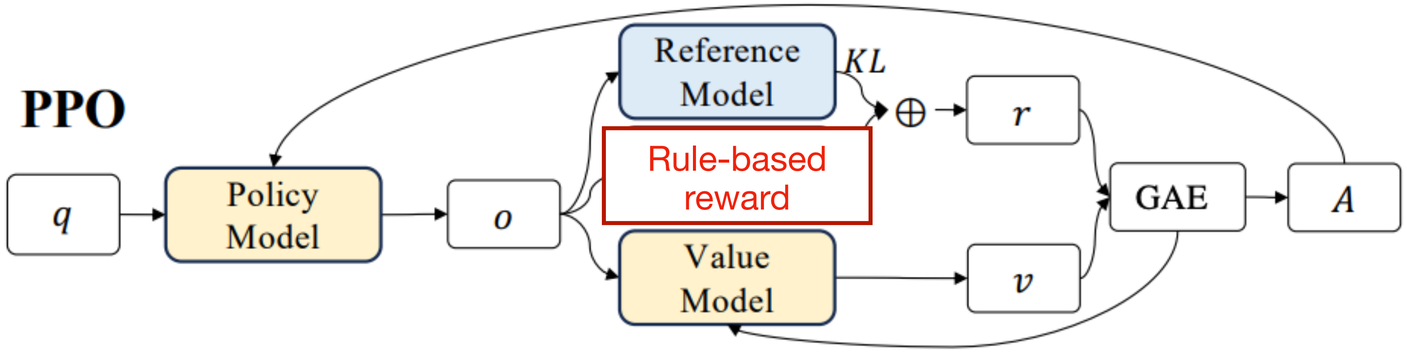
We use the PPO algorithm, the training objective is:

$$\mathcal{L}_{\text{ppo-clip}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right],$$



Shao Z, Wang P, Zhu Q, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models[J]. arXiv preprint arXiv:2402.03300, 2024.

The SimpleRL Recipe



- Rule-based reward function (Correctness and Format)
  - If the answer is correct, reward = 1
  - If the format is correct but the answer is incorrect, reward = -0.5
  - If the format is incorrect, reward = -1

Format Example

Many many reasoning steps….

Since \(3^{1/7}\) is the seventh root of 3, the minimum value is: \boxed{36}.

In the following experiments, we only use 8K examples from MATH dataset.

Experiment Setup

- Start from the Qwen2.5-Math-7B-Base model.

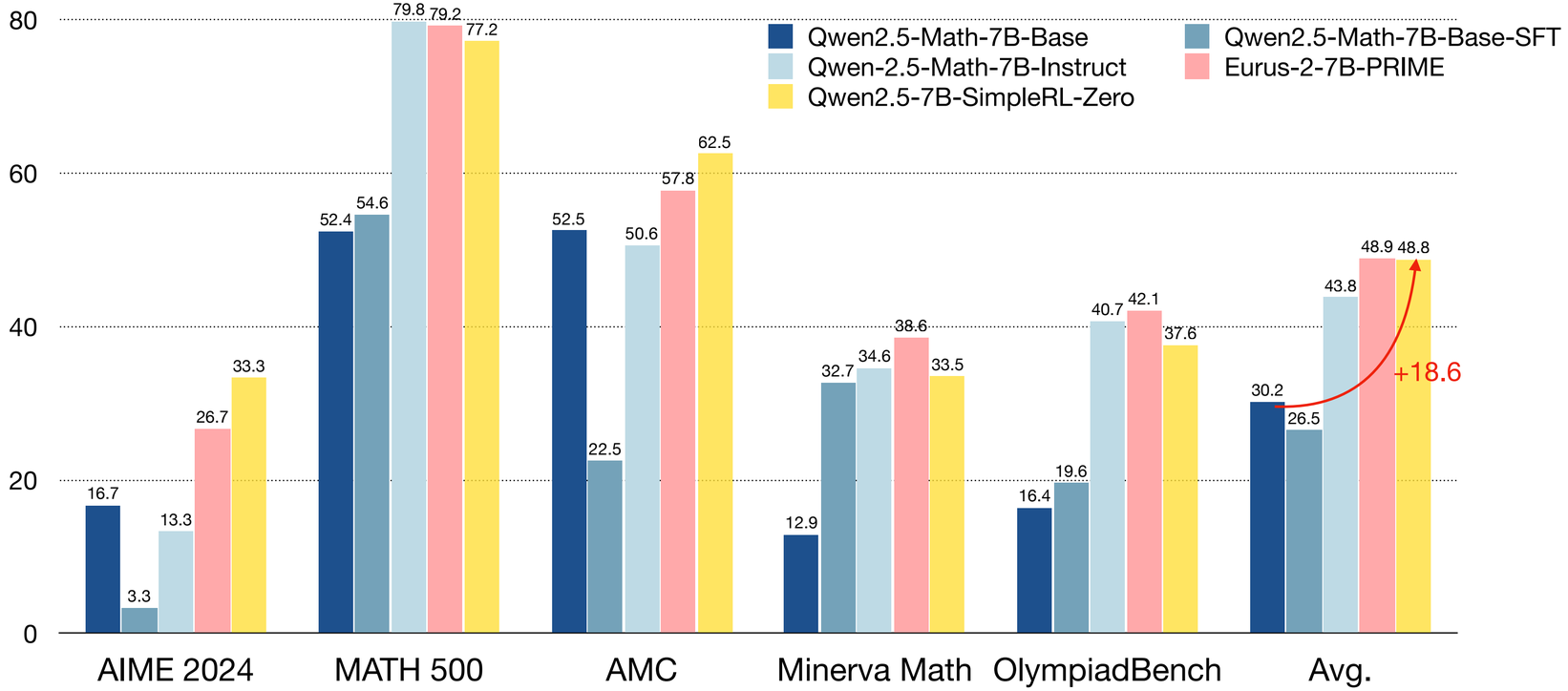
	Qwen2.5-Math-7B-Instruct	rStar-Math-7B	Eurus-2-7B-PRIME	Qwen2.5-7B-SimpleRL-Zero
Base Model	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B	Qwen2.5-Math-7B
SFT Data	2.5M (open-source and in-house)	~7.3 M (MATH, NuminaMath, etc.)	230K	0
RM Data	618K (in-house)	~7 k (in-house)	0	0
RM	Qwen2.5-Math-RM (72B)	None	Eurus-2-7B-SFT	None
RL Data	66K queries × 32 samples	~3.647 M × 16	150K queries × 4 samples	8K queries × 8 samples

Using only 8K examples from the original MATH dataset, much less than other methods.

Evaluate on challenging math benchmarks, AIME, AMC …

SimpleRL-Zero — RL from scratch

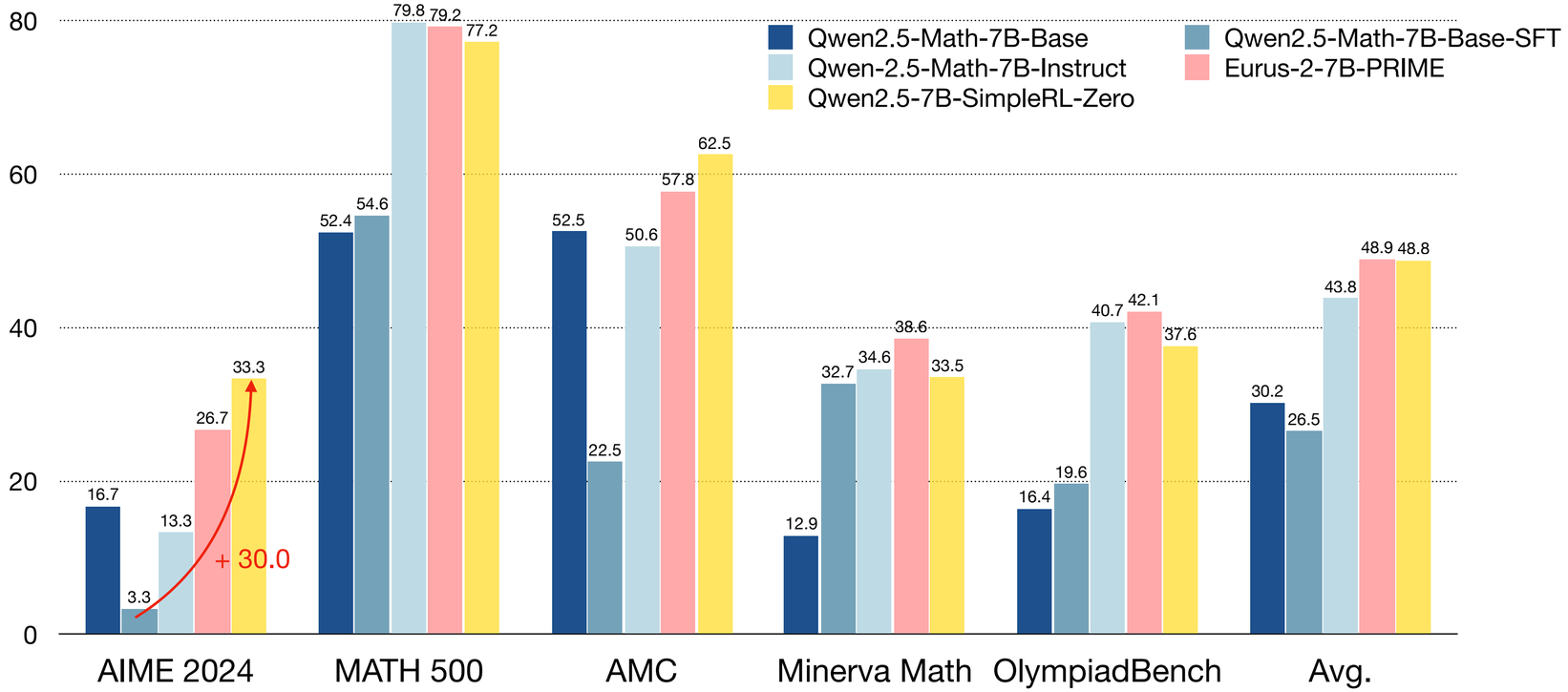
RL is applied directly to the base model using 8K MATH (query, answer) pairs without SFT.



Achieve gains of nearly 20 absolute points from base model.

Outperform Instruct model and achieves comparable results to PRIME with 50x data efficiency.

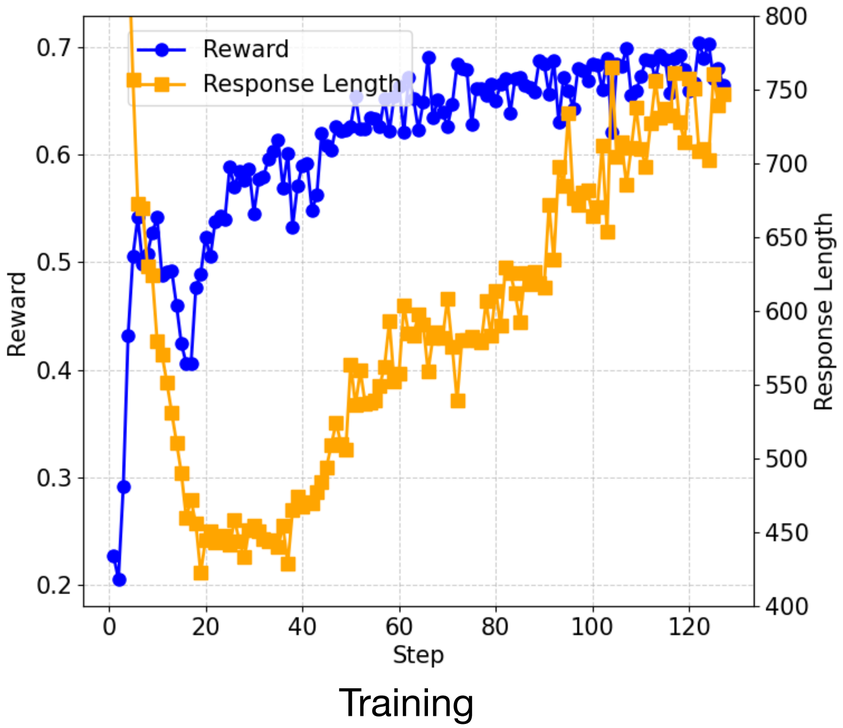
SimpleRL-Zero — RL from scratch



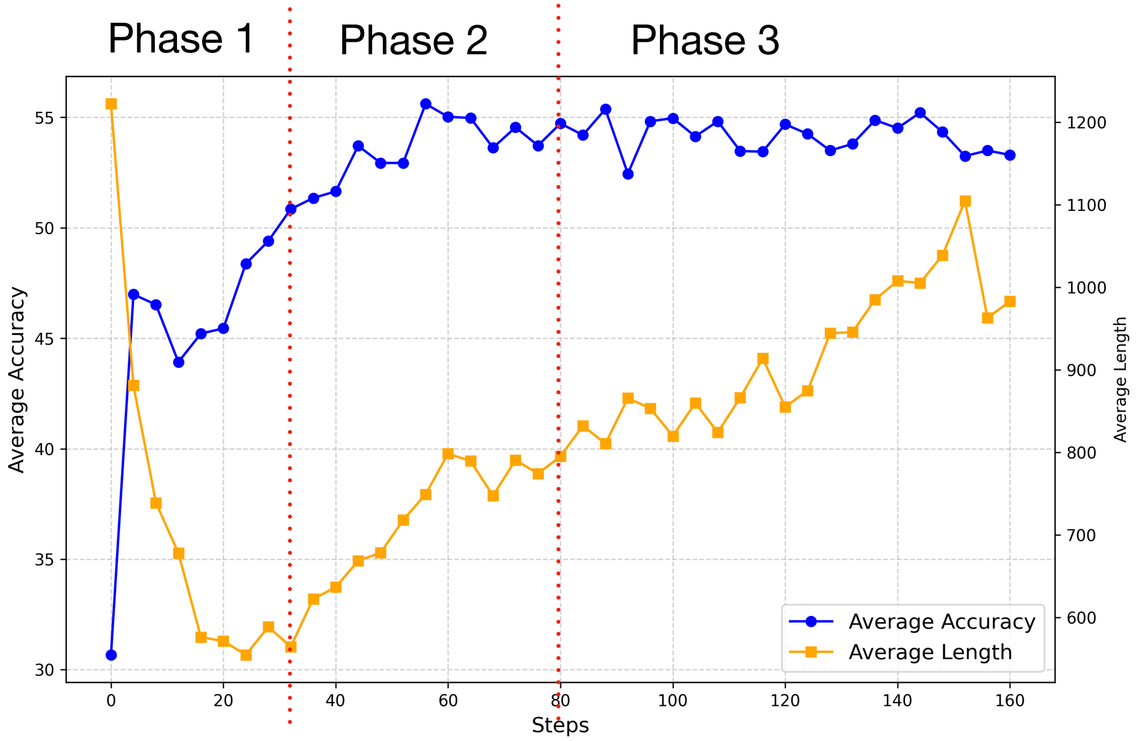
Compared to SFT, RL generalizes to competition-level benchmarks and being 30 points higher in AIME

Easy-to-hard generalization !

SimpleRL-Zero — RL from scratch



Training



Evaluation

The accuracy is improving on average, while the length decreases first and then gradually increases.

Surprisingly how far the 8K MATH examples lift this 7B base model without any other external signals.

Almost all open reproduction of DeepSeek-R1 is based Qwen models, but Qwen models are kinda special.

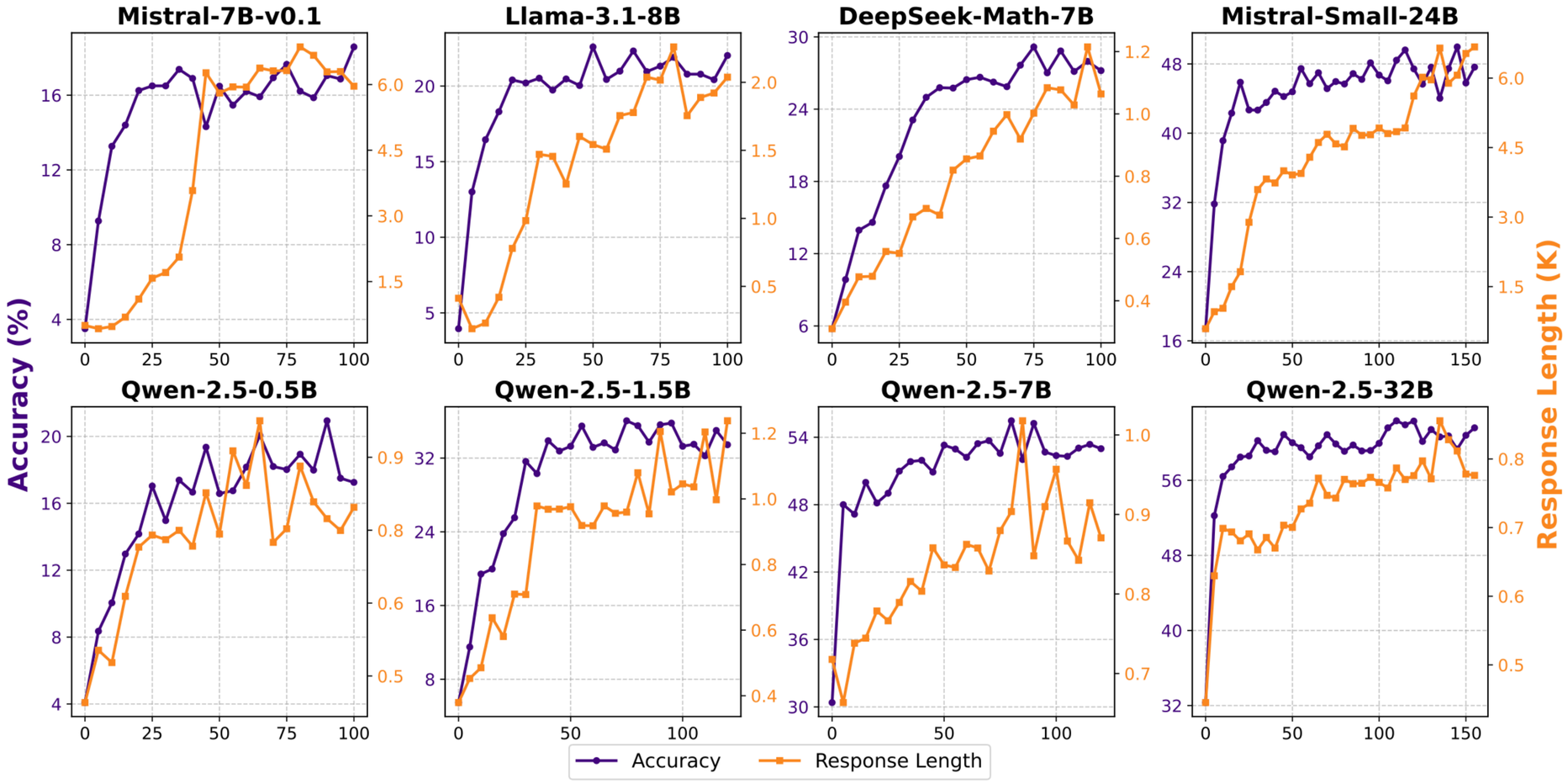
Are the conclusions from these works Qwen-specific?

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

SimpleRL-Zoo: Investigating and Taming Zero Reinforcement Learning for Open Base Models in the Wild

Weihao Zeng*, Yuzhen Huang*, Qian Liu*, Wei Liu, Keqing He, Zejun Ma, Junxian He

Consistent Results



Main Results

Model	GSM8K	MATH 500	Minerva Math	Olympiad Bench	AIME24 (Pass@1)	AIME24 (Avg@32)	AMC23	Avg.
Llama, DeepSeek and Mistral Models								
Mistral-v0.1-7B	21.2	4.2	4.0	2.4	0.0	0.0	0.0	5.3
↪ + SimpleRL-Zoo	75.0	15.8	6.6	4.1	0.0	0.2	10.0	18.6
Llama-3.1-8B	39.7	13.6	4.8	3.1	0.0	0.2	2.5	10.6
↪ + SimpleRL-Zoo	79.2	23.0	9.6	5.3	0.0	0.2	15.0	22.0
DeepSeek-Math-7B	28.4	19.4	5.5	4.7	0.0	0.0	10.0	11.3
↪ + SimpleRL-Zoo	78.5	39.6	21.0	12.6	3.3	0.6	20.0	29.2
Mistral-Small-24B	78.6	43.6	10.7	11.6	3.3	0.5	17.5	27.6
↪ + SimpleRL-Zoo	92.0	70.6	36.8	36.6	16.7	13.1	45.0	49.6
Qwen Series Models								
Qwen-2.5-0.5B	36.7	15.8	4.8	2.8	0.0	0.3	12.5	12.1
↪ + SimpleRL-Zoo	49.5	34.4	10.3	8.9	0.0	0.7	22.5	20.9
Qwen-2.5-1.5B	55.7	29.6	6.6	6.5	0.0	0.1	12.5	18.5
↪ + SimpleRL-Zoo	74.4	59.0	20.2	21.0	6.7	4.2	35.0	36.1
Qwen-2.5-7B	88.2	64.6	25.7	30.1	3.3	0.3	30.0	40.3
↪ + SimpleRL-Zoo	91.7	78.2	38.6	40.4	20.0	15.6	62.5	55.2
Qwen-2.5-Math-7B	65.5	63.6	12.5	25.8	13.3	8.6	42.5	37.2
↪ + SimpleRL-Zoo	90.2	80.2	37.5	39.0	40.0	24.0	70.0	59.5
Qwen-2.5-14B	91.6	65.4	24.3	33.5	6.7	3.4	37.5	43.2
↪ + SimpleRL-Zoo	94.4	80.2	40.4	44.9	23.3	14.2	57.6	56.8
Qwen-2.5-32B	92.9	68.6	27.9	31.1	10.0	4.5	45.0	45.9
↪ + SimpleRL-Zoo	95.9	82.4	42.6	46.4	36.7	27.2	67.5	61.9

RL Generalizes

Model	IFEVAL strict-prompt	MMLU Stem	MMLU	GPQA	Avg.
Llama, DeepSeek and Mistral Models					
Mistral-v0.1-7B	13.5	26.1	28.0	23.2	22.7
↪ + SimpleRL-Zoo	21.8	28.1	34.6	30.3	28.7
Llama-3.1-8B	16.1	27.1	28.7	22.7	23.6
↪ + SimpleRL-Zoo	25.1	40.7	44.5	20.2	32.6
DeepSeek-Math-7B	11.5	21.6	22.7	19.2	18.7
↪ + SimpleRL-Zoo	16.3	47.4	45.5	27.3	34.1
Mistral-Small-24B	17.4	30.9	31.7	20.2	25.0
↪ + SimpleRL-Zoo	23.5	73.9	78.8	45.0	55.3
Qwen Series Models					
Qwen-2.5-0.5B	9.6	23.2	24.9	24.8	20.6
↪ + SimpleRL-Zoo	14.4	32.1	34.6	26.3	26.8
Qwen-2.5-1.5B	15.2	33.1	35.4	24.8	27.1
↪ + SimpleRL-Zoo	20.3	42.1	45.2	28.8	34.1
Qwen-2.5-7B	21.3	39.8	38.6	23.7	30.8
↪ + SimpleRL-Zoo	25.9	49.6	47.0	29.8	38.1
Qwen-2.5-Math-7B	14.1	40.6	38.0	27.8	30.1
↪ + SimpleRL-Zoo	17.0	55.6	56.6	35.4	41.1
Qwen-2.5-14B	22.9	59.8	63.5	24.8	42.7
↪ + SimpleRL-Zoo	29.4	76.3	79.1	50.0	58.7
Qwen-2.5-32B	24.6	60.7	62.7	38.9	46.7
↪ + SimpleRL-Zoo	31.2	79.0	82.5	49.5	60.6

Revisit the Old Doubt on RL

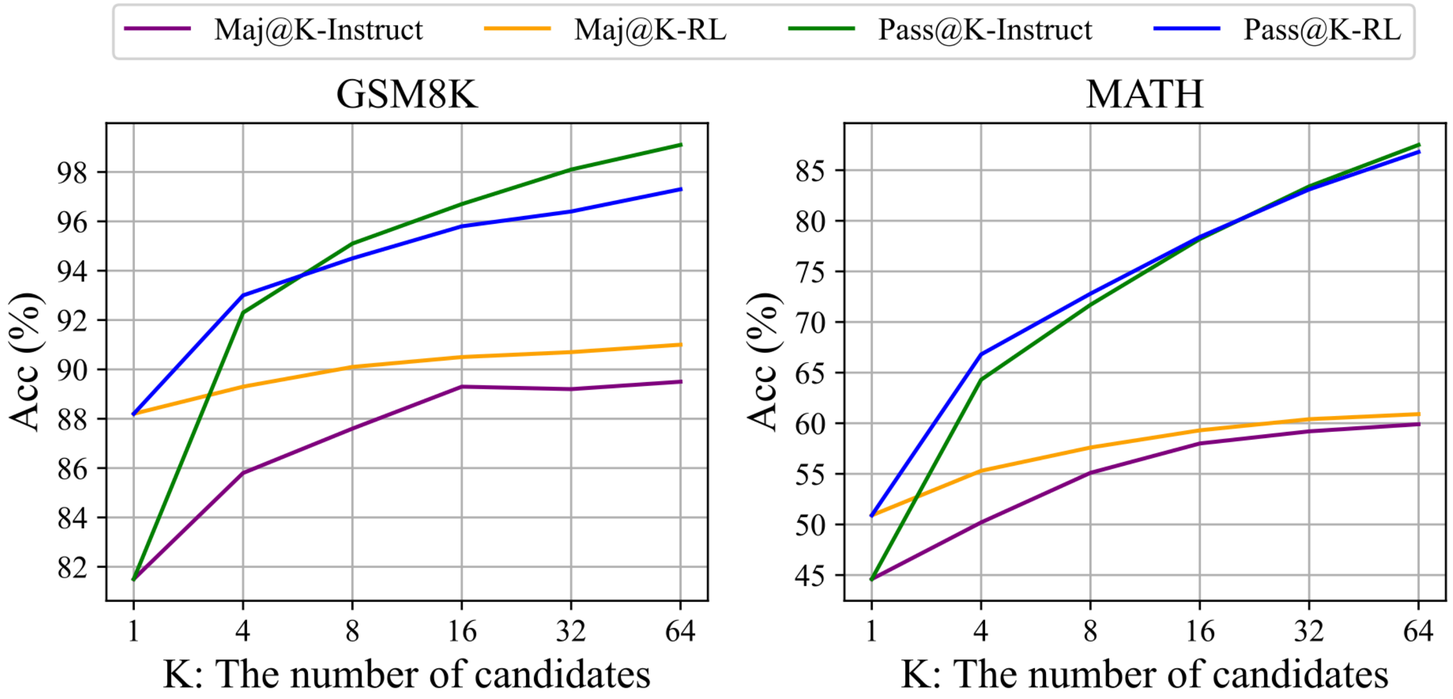


Figure 7 | The Maj@K and Pass@K of SFT and RL DeepSeekMath 7B on GSM8K and MATH (temperature 0.7). It was noted that RL enhances Maj@K but not Pass@K.

For a long time, a lot of people don't believe RL can fundamentally improve model abilities

[1] Shao et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. 2024

But it is different now

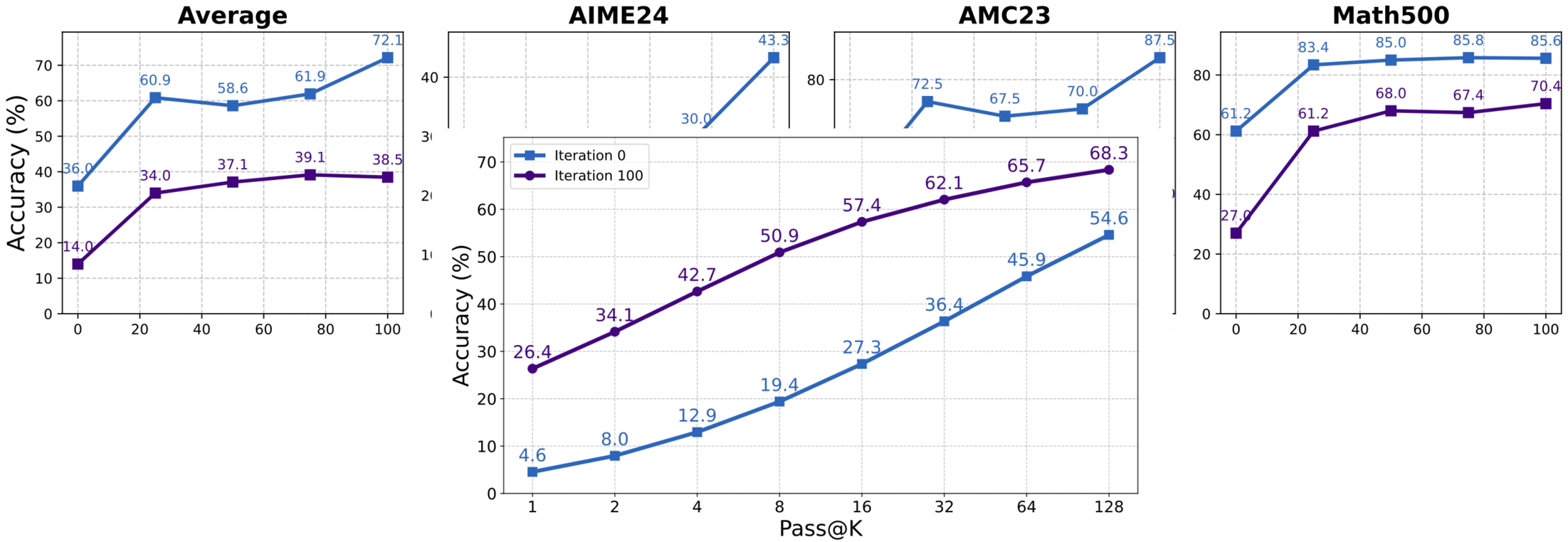
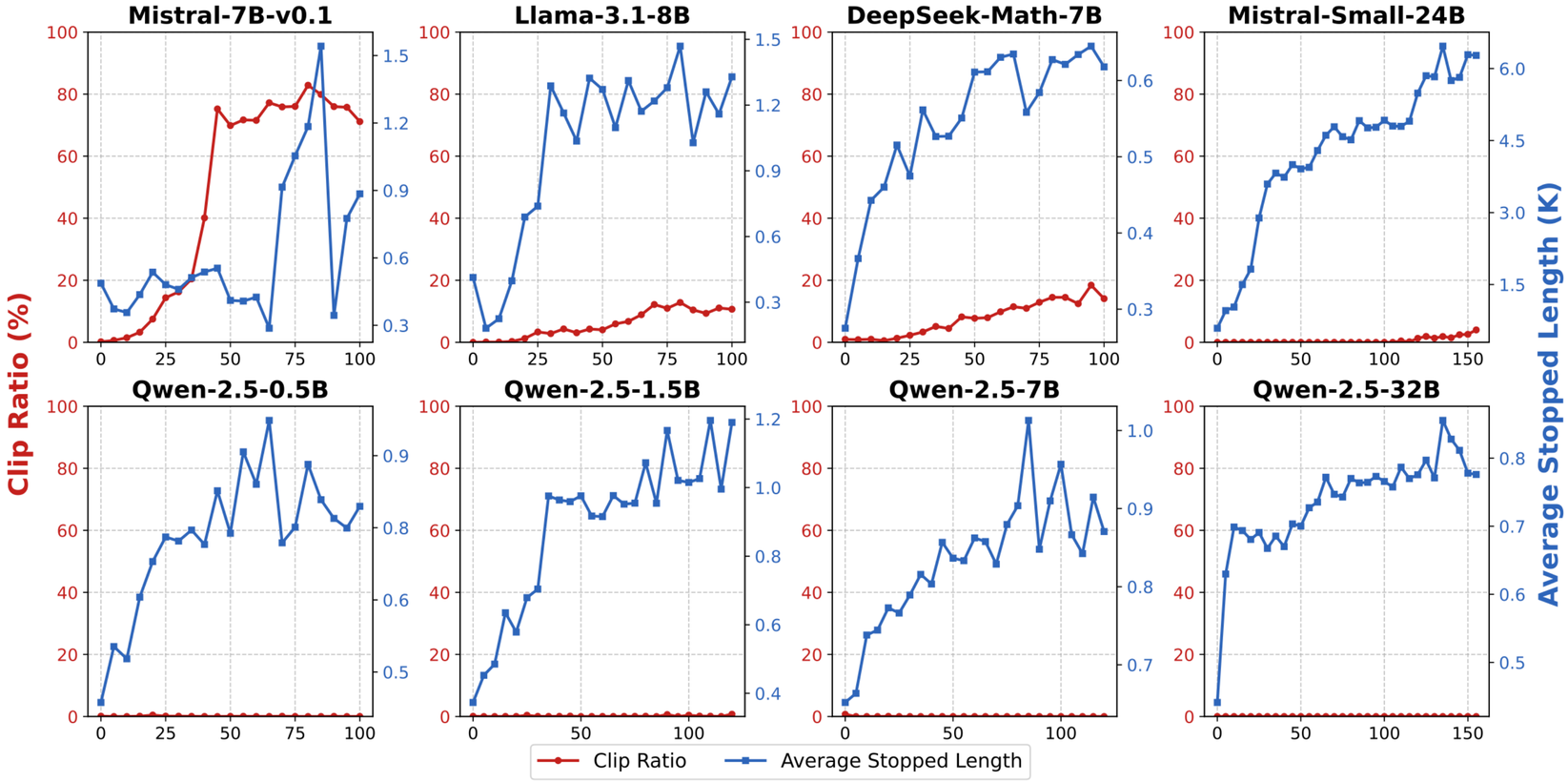


Figure 3: Pass@k of Mistral-24B based on the average results from AIME24 and AMC23.

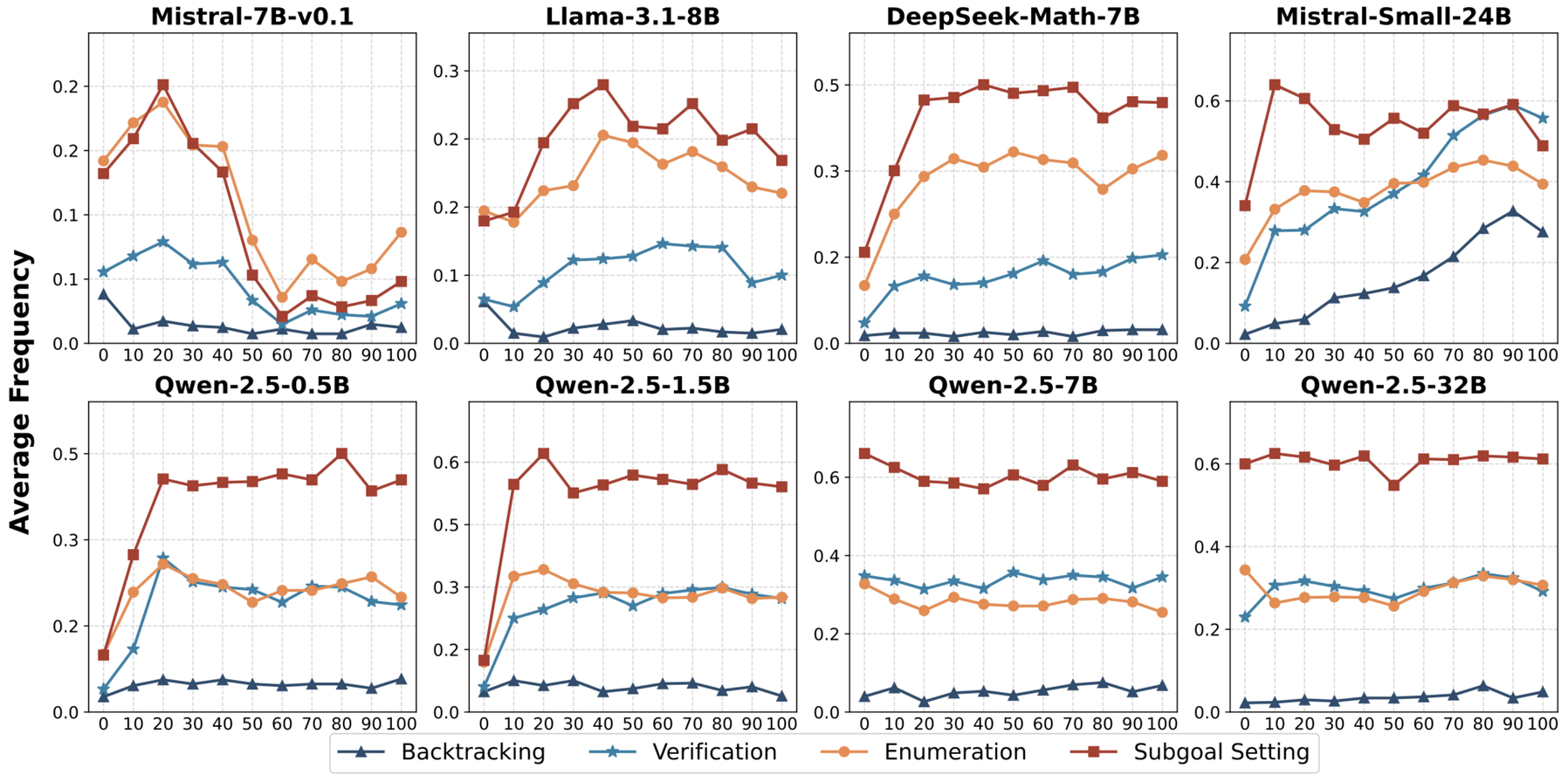
Model's reasoning abilities are fundamentally improved

Response Length does not reflect the full story



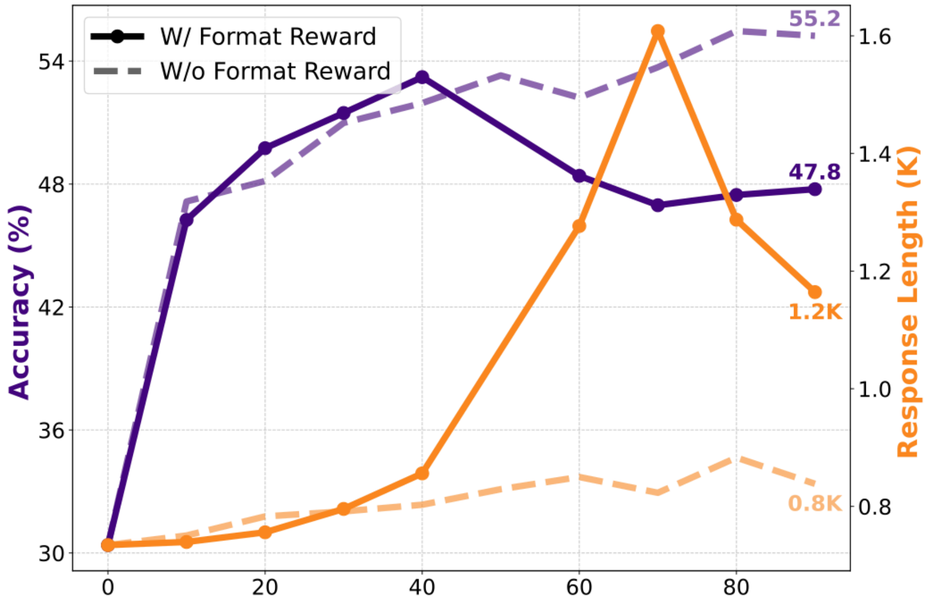
Response length increase of Mistral-7B is unhealthy

Cognitive Behavior

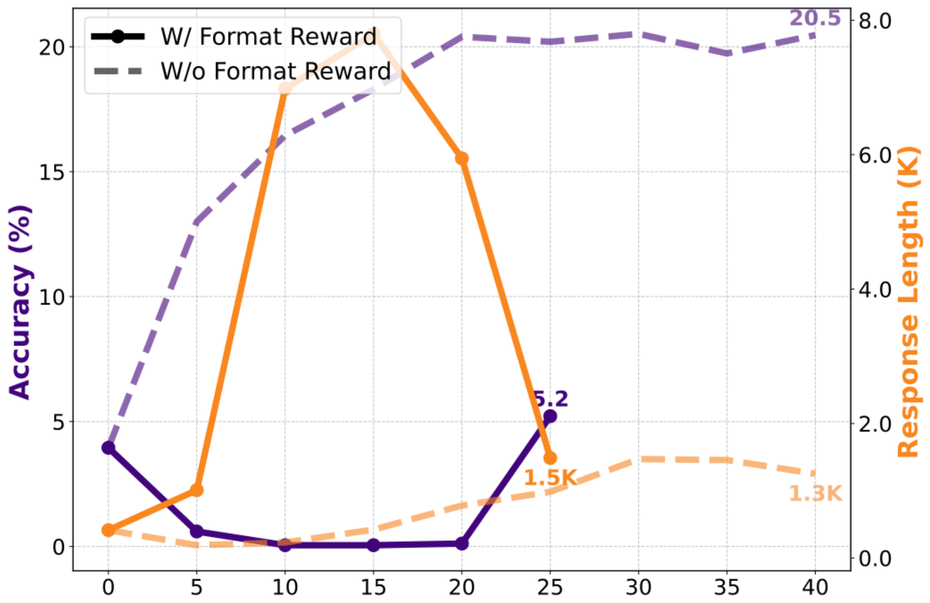


Response length is just a superficial metric, we should focus more on the true metric

Format Reward is Not Always a Good Thing



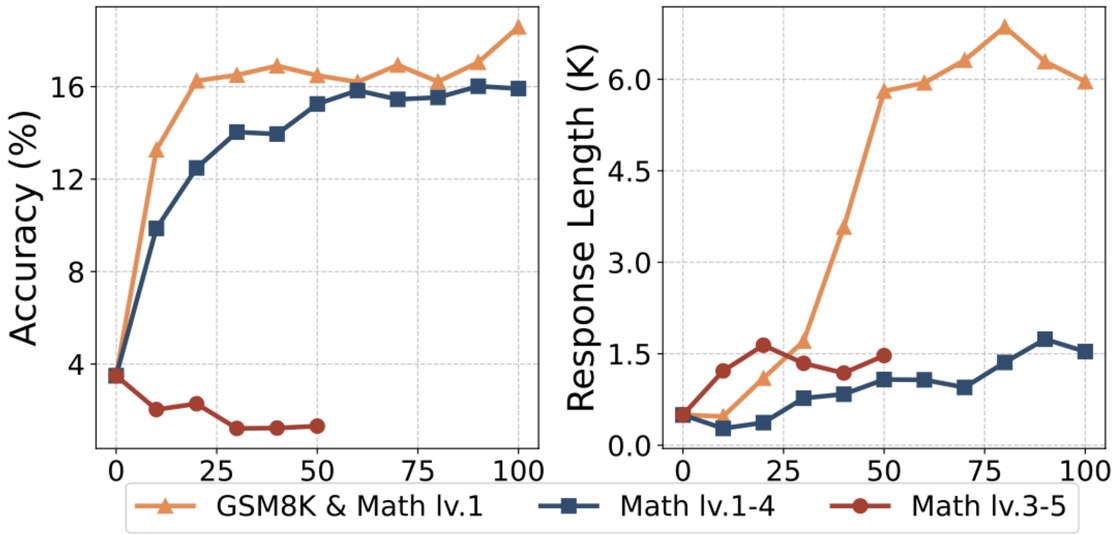
(a) Qwen-2.5-7B



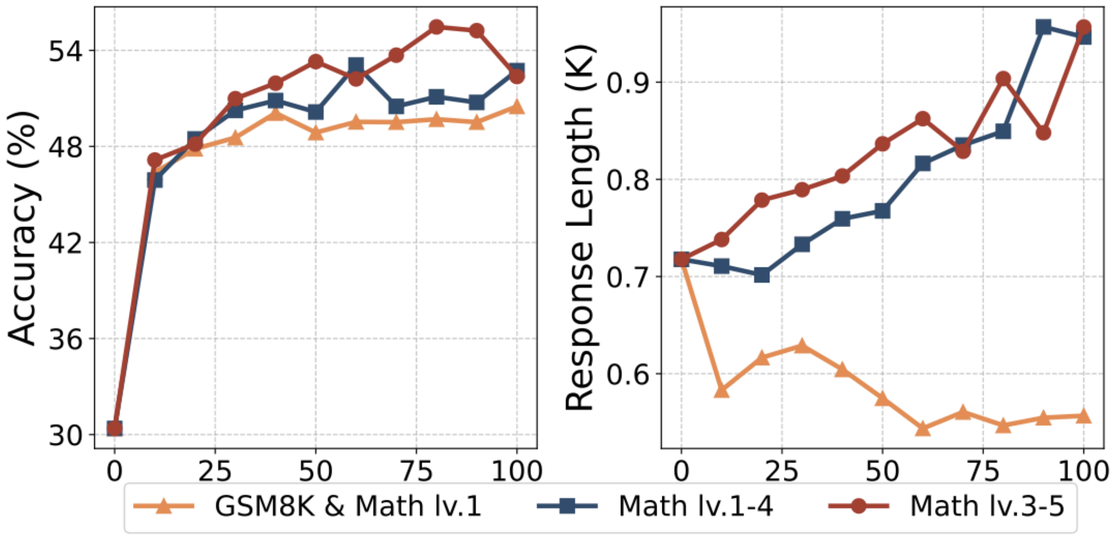
(b) Llama-3.1-8B

Applying format reward limits exploration in the initial stage and hurts particularly for weaker base models

Data Difficulty, not too hard, not too easy



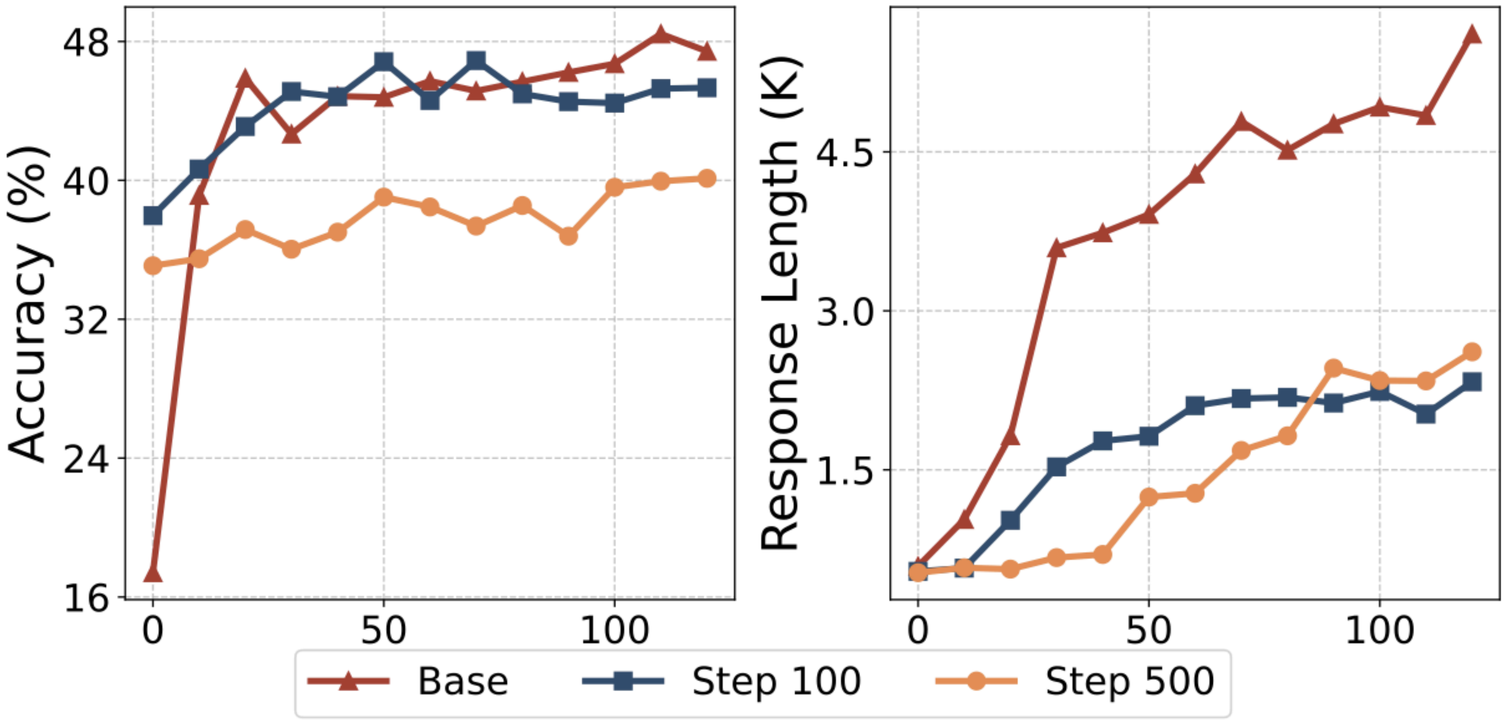
(a) Mistral-7b-v0.1



(b) Qwen-2.5-7B

Is Cold-Start Always a Good Thing for RL?

Revisit the pretraining -> short CoT SFT -> RL recipe Why RL fails in the past?



The worse it gets when we perform more SFT before RL

Is Cold-Start Always a Good Thing for RL?

Revisit the pretraining -> short CoT SFT -> RL recipe Why RL fails in the past?

Init Model	GSM8K	MATH 500	Minerva Math	Olympiad Bench	AIME24 (pass@1)	AMC23	Avg.
0 Step	92.0	70.6	36.8	36.6	16.7	45.0	49.6
10 Step	93.0	69.4	39.7	32.3	10.4	44.1	48.2
20 Step	92.6	65.2	34.2	30.7	6.7	38.4	44.6
200 Step	90.3	59.0	31.6	23.3	2.1	26.9	38.9
1000 Step	88.9	48.8	27.6	20.7	2.5	18.1	34.4
2000 Step	89.8	49.0	23.2	18.1	0.8	20.3	33.5
4000 Step	87.7	52.0	23.5	17.2	2.1	21.6	34.0

The worse it gets when we perform more SFT before RL

Outline

1. Introduction

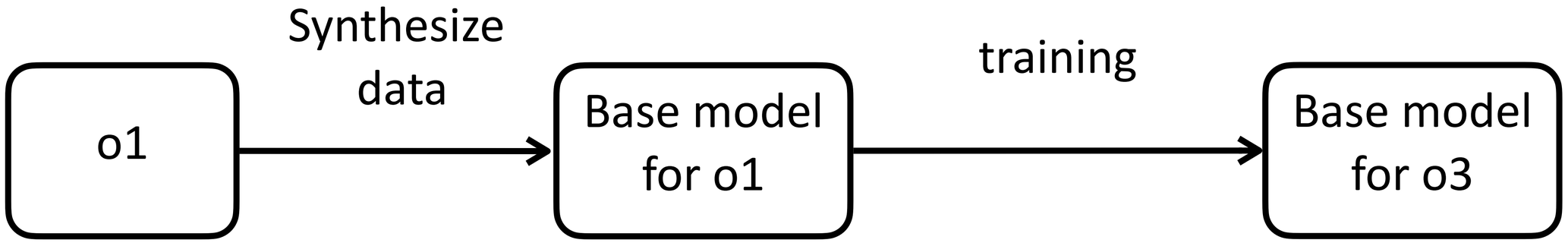
2. How to define good synthetic data in self-improving?

3. SimpleRL-Zoo: RL for open base models in the wild

4. Future Directions

Future Directions

GPT4 -> o1, how o1 -> o3?



Future Directions

1. Reward for various environments

Direct seeking of rewards may be impossible for scalable RL in some tasks

2. Native fast and slow thinking

Can we bypass long CoT to achieve scalable RL in LLM?

3. CoT faithfulness

4. Unnatural reasoning chains

Thank You!