Reinforcement Learning for Reasoning in Large Language Models with ***One*** Training Example

Yiping Wang

University of Washington & Microsoft

Reinforcement Learning with Verifiable Reward

- Previously, RLHF is widely used for the post-training stage of large language models (LLMs)

- They always use a trained (process/outcome) reward model for providing reward signal in RLHF, which may suffer from several issues:
  - Reward hacking
  - Accuracy of the reward model is not that high
  - High training cost

Reinforcement Learning with Verifiable Reward

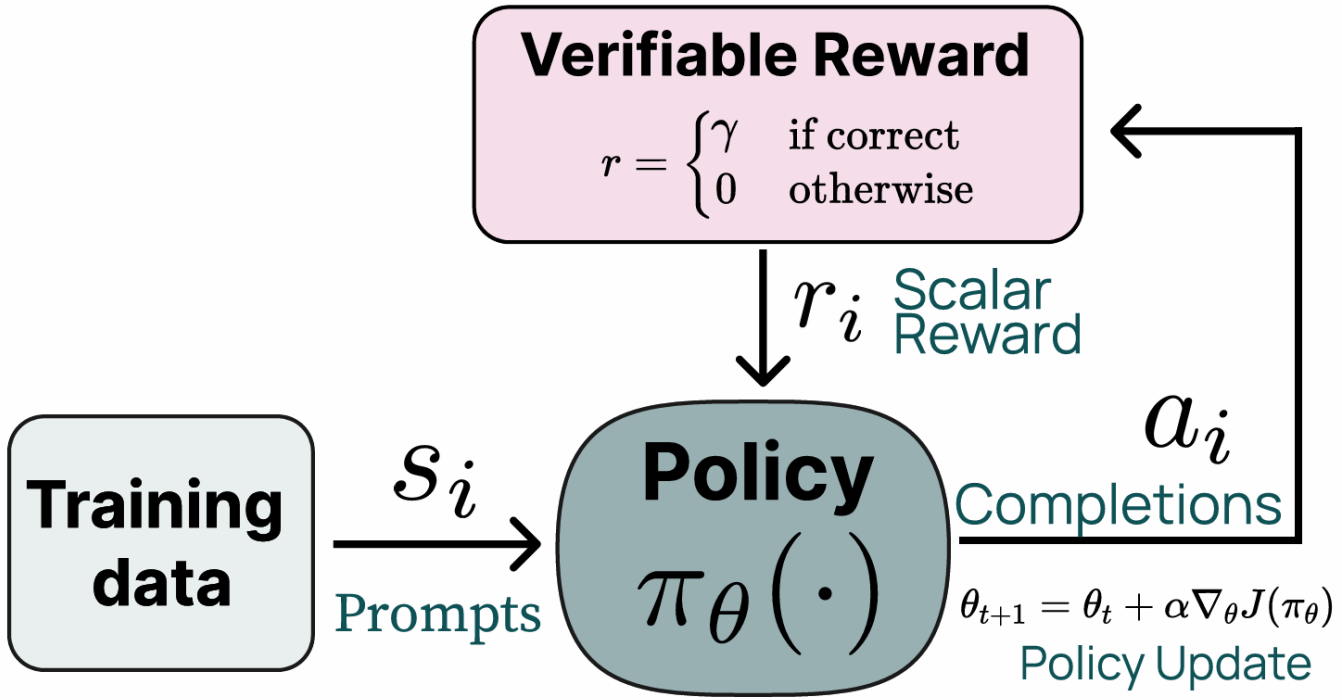
- Previously, RLHF is widely used for the post-training stage of large language models (LLMs)

- They always use a trained (process/outcome) reward model for providing reward signal in RLHF, which may suffer from several issues:
  - Reward hacking
  - Accuracy of the reward model is not that high
  - High training cost

- **RLVR** (Reinforcement Learning with Verifiable Reward) is becoming more popular nowadays for training reasoning LM. It only requires a rule-based outcome reward

Reinforcement Learning with Verifiable Reward

- **RLVR** is widely used in improving LLM performance in reasoning tasks (math, code, etc.)
- Use verifiable outcome reward in RL training (e.g. 0-1 correctness reward for math data)



Lambert, Nathan, et al. "Tulu 3: Pushing Frontiers in Open Language Model Post-Training" arXiv preprint arXiv:2411.15124 (2024).

Reinforcement Learning with Verifiable Reward

- Used in advanced reasoning LLM like DeepSeek-R1, kimi-1.5, etc.
- Combined with RL algorithms like PPO and **GRPO**.

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1,$$

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}.$$

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

Reinforcement Learning with Verifiable Reward

GRPO:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1,$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}.$$

- q: question from dataset.
- o_i: i-th generated outputs from old policy model θ_{old}
- G: group size
- ϵ: clipping hyperparameter
- r_i: reward for o_i
- A_i: group advantage for o_i
- D_{KL}: KL divergence between current policy θ and reference policy θ_{ref}.

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

Reinforcement Learning with Verifiable Reward

- Many recent works focus on designing new RL algorithms:

  REINFORCE++, VinePPO, VC-PPO, VAPO, DAPO, Dr. GRPO, GRPO+, SRPO, EMPO, …

- It's relatively underexplored in how **data** affects RLVR

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^{G}|o_i|} \sum_{i=1}^{G}\sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\right)\hat{A}_{i,t}\right)\right]$$
$$\text{s.t.} \quad 0 < \left|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\right| < G.$$

Dr. GRPO

GRPO Done Right (without bias)

$$\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|\mathbf{o}_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|\mathbf{q},\mathbf{o}_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right]\right\},$$

$$\text{where } \hat{A}_{i,t} = R(\mathbf{q},\mathbf{o}_i) - \text{mean}(\{R(\mathbf{q},\mathbf{o}_1),\ldots,R(\mathbf{q},\mathbf{o}_G)\}).$$

Yu, Qiying, et al. "Dapo: An open-source llm reinforcement learning system at scale." arXiv preprint arXiv:2503.14476 (2025).
Liu, Zichen, et al. "Understanding r1-zero-like training: A critical perspective." arXiv preprint arXiv:2503.20783 (2025).

Data Selection in RLVR

Q: To what extent can we ***reduce*** the training dataset for RLVR while maintaining comparable performance compared to using the full dataset?

Data Selection in RLVR

Q: To what extent can we ***reduce*** the training dataset for RLVR while maintaining comparable performance compared to using the full dataset?

ONE

Evaluation Dataset

Mathematical reasoning tasks:

- MATH500
- AIME2024
- AIME2025
- AMC2023
- Minerva Math
- OlympiadBench

Non-mathematical reasoning tasks:

- ARC-Easy/Challenge

Evaluation Dataset

Mathematical reasoning tasks:

- MATH500
- AIME2024
- AIME2025
- AMC2023
- Minerva Math
- OlympiadBench

Non-mathematical reasoning tasks:

- ARC-Easy/Challenge

Example from MATH500:

```
Convert the point $(0,3)$ in
rectangular coordinates to polar
coordinates. Enter your answer in the
form $(r,\theta),$ where $r > 0$ and
$0 \le \theta < 2 \pi.$
```

Evaluation Dataset

Mathematical reasoning tasks:

- MATH500
- AIME2024
- AIME2025
- AMC2023
- Minerva Math
- OlympiadBench

Non-mathematical reasoning tasks:

- ARC-Easy/Challenge

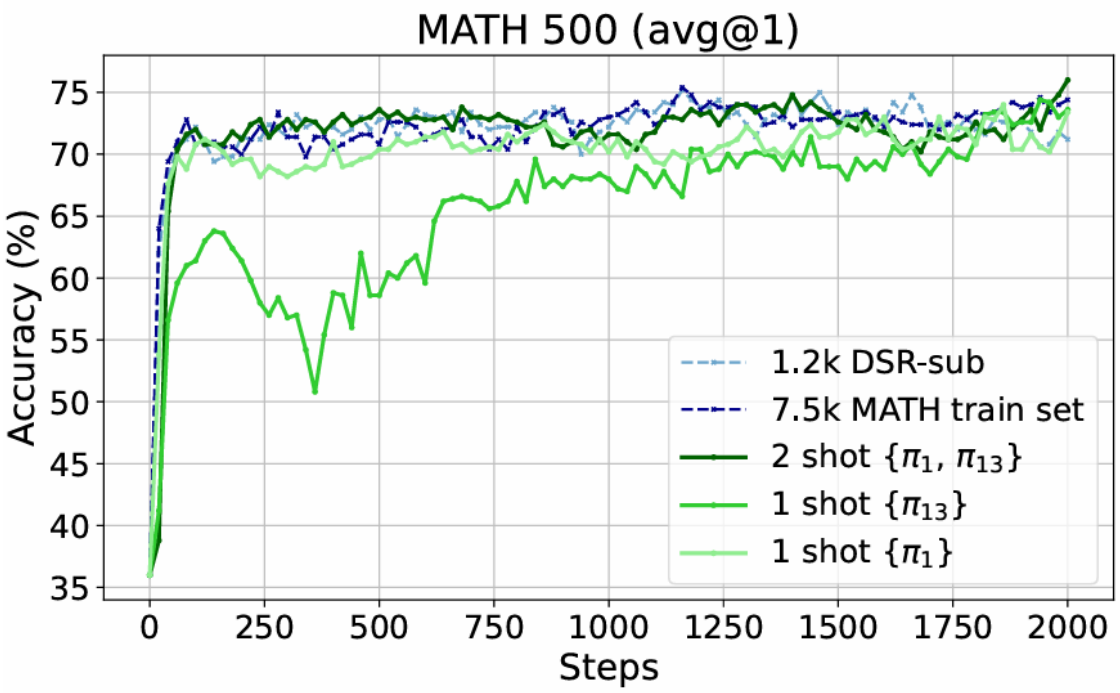
Example from ARC-Challenge:

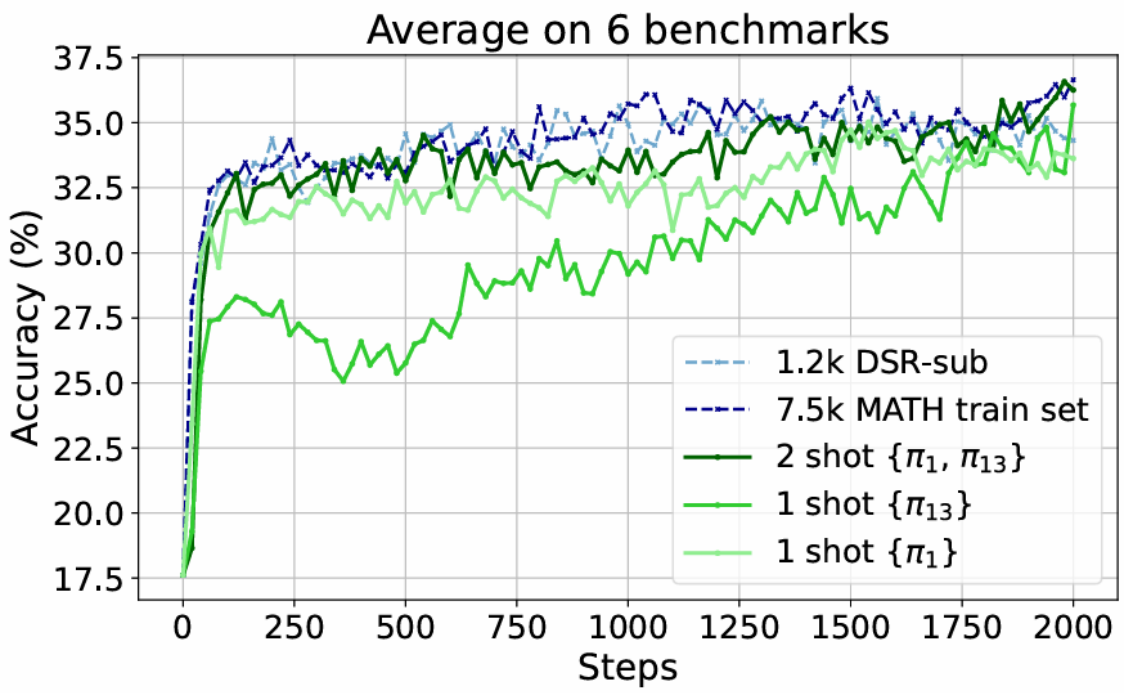
George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?

A. Dry palms
B. Wet palms
C. Palms covered with oil
D. Palms covered with lotion

One-Shot RLVR

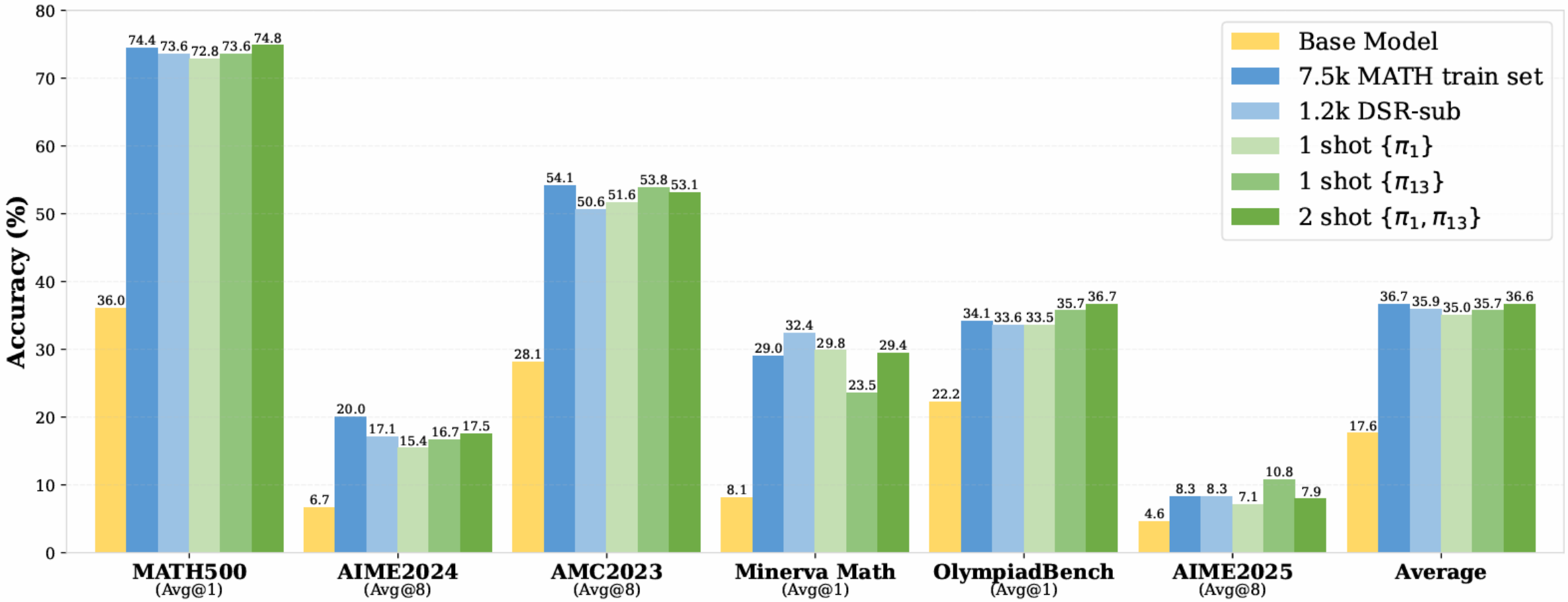
- **Model:** Qwen2.5-Math-1.5B, **Data Pool**: 1.2k DeepScaleR-subset (DSR-sub)
- 1-shot RLVR works as well as 1.2k DSR-sub dataset (which contain that one example)





One-Shot RLVR

- Improves a lot compared from base model on 6 math reasoning benchmarks



One-Shot RLVR

- 1-shot RLVR with math example can even improve model performance on non-math tasks (ARC-Easy/Challenge), even better than full-set RLVR.

Dataset	Size	ARC-E	ARC-C
Base	NA	48.0	30.2
MATH	7500	51.6	<u>32.8</u>
DSR-sub	1209	42.2	29.9
$\{\pi_1\}$	1	52.0	32.2
$\{\pi_{13}\}$	1	**55.8**	**33.4**
$\{\pi_1, \pi_{13}\}$	2	<u>52.1</u>	32.4

RLVR Loss

- We follow the default setup of **verl**, which include three losses by default
  - **Policy gradient loss**: normal GRPO loss
  - **KL divergence loss** ($\beta > 0$)
  - **Entropy loss** ($\alpha < 0$): per-token entropy, for encouraging exploration

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q \sim P(Q) \\ \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)}} \left[\mathcal{L}'_{\text{PG-GRPO}}(\cdot, \theta) + \beta \mathcal{L}'_{\text{KL}}(\cdot, \theta, \theta_{\text{ref}}) + \alpha \mathcal{L}'_{\text{Entropy}}(\cdot, \theta) \right]$$

Sheng, Guangming, et al. "Hybridflow: A flexible and efficient rlhf framework." arXiv preprint arXiv:2409.19256 (2024).

Data Selection: Historical Variance Score

Motivation: Previous work shows that

- the variance of the reward signal is critical for RL training [1]
- choosing problems with medium difficulty will be better [2, 3]

We design a score named historical variance score to rank the data

[1] Razin, Noam, et al. "What makes a reward model a good teacher? an optimization perspective." arXiv preprint arXiv:2503.15477 (2025).
[2] Yu, Qiying, et al. "Dapo: An open-source llm reinforcement learning system at scale." arXiv preprint arXiv:2503.14476 (2025).
[3] Li, Xuefeng, Haoyang Zou, and Pengfei Liu. "Limr: Less is more for rl scaling." arXiv preprint arXiv:2502.11886 (2025).

Data Selection: Historical Variance Score

(1) Train RLVR for E epochs on full dataset, obtain historical training acc for each example i.

(2) Rank the data by their historical variance of acc.

$$L_i = [s_{i,1}, \ldots, s_{i,E}]$$

$$v_i := \mathrm{var}(s_{i,1}, \ldots, s_{i,E})$$

$$\pi_j := \pi(j) = \arg\operatorname*{sort}_j \{v_i : i \in [N]\}$$

Data Selection: Historical Variance Score

(1) Train RLVR for E epochs on full dataset, obtain historical training acc for each example i.

$$L_i = [s_{i,1}, \ldots, s_{i,E}]$$

(2) Rank the data by their historical variance of acc.

$$v_i := \mathrm{var}(s_{i,1}, \ldots, s_{i,E})$$

This criterion is **not necessarily optimal**! 1-shot RLVR works for a lot of examples.

$$\pi_j := \pi(j) = \arg\operatorname*{sort}_j \{v_i : i \in [N]\}$$

Data Selection: Historical Variance Score

- We copy the single example many times to fill the entire training batch (e.g. 128)
- (just because verl requires at least one example allocated to each GPU)

$$L_i = [s_{i,1}, \ldots, s_{i,E}]$$

$$v_i := \mathrm{var}(s_{i,1}, \ldots, s_{i,E})$$

$$\pi_j := \pi(j) = \arg\operatorname*{sort}_j \{v_i : i \in [N]\}$$

Dissection of Selected Examples

- Not-so-difficult problems: initial model is already capable of sampling correct answers

Prompt of example π_1:

```
The pressure \\( P \\) exerted by wind on a sail varies jointly as the area \\( A \\) of the sail and the cube of the wind's velocity \\( V \\). When the velocity is \\( 8 \\) miles per hour, the pressure on a sail of \\( 2 \\) square feet is \\( 4 \\) pounds. Find the wind velocity when the pressure on \\( 4 \\) square feet of sail is \\( 32 \\) pounds. Let's think step by step and output the final answer within \\boxed{}.
```

Ground truth (label in DSR-sub): 12.8.

Prompt of example π_{13}:

```
Given that circle $C$ passes through points $P(0,-4)$, $Q(2,0)$, and $R(3,-1)$.  \n$(1)$ Find the equation of circle $C$.  \n$(2)$ If the line $l: mx+y-1=0$ intersects circle $C$ at points $A$ and $B$, and $|AB|=4$, find the value of $m$. Let's think step by step and output the final answer within \\boxed{}.
```

Ground truth (label in DSR-sub): $\frac{4}{3}$.

A Universal Phenomenon

- Almost all examples can be used in 1-shot RLVR

Dataset	Size	Step	Type	Alg.	C. P.	Geo.	I. Alg.	N. T.	Prealg.	Precal.	MATH500	AIME24
Base	0	0	NA	37.1	31.6	39.0	43.3	24.2	36.6	33.9	36.0	6.7
MATH	7500	1160	General	91.1	65.8	63.4	59.8	82.3	81.7	66.1	75.4	**20.4**
DSR-sub	1209	1160	General	91.9	68.4	58.5	57.7	**85.5**	79.3	67.9	75.2	18.8
$\{\pi_1\}$	1	1860	Alg.	88.7	63.2	56.1	**62.9**	79.0	81.7	64.3	74.0	16.7
$\{\pi_2\}$	1	220	N. T.	83.9	57.9	56.1	55.7	77.4	82.9	60.7	70.6	17.1
$\{\pi_4\}$	1	80	N. T.	79.8	57.9	53.7	51.6	71.0	74.4	53.6	65.6	17.1
$\{\pi_7\}$	1	580	I. Alg.	75.8	60.5	51.2	56.7	59.7	70.7	57.1	64.0	12.1
$\{\pi_{11}\}$	1	20	N. T.	75.8	65.8	56.1	50.5	66.1	73.2	50.0	64.0	13.3
$\{\pi_{13}\}$	1	1940	Geo.	89.5	65.8	63.4	55.7	83.9	81.7	66.1	74.4	17.1
$\{\pi_{16}\}$	1	600	Alg.	86.3	63.2	56.1	51.6	67.7	73.2	51.8	67.0	14.6
$\{\pi_{17}\}$	1	220	C. P.	80.7	65.8	51.2	58.8	67.7	78.1	48.2	67.2	13.3
$\{\pi_{605}\}$	1	1040	Precal.	84.7	63.2	58.5	49.5	82.3	78.1	62.5	71.8	14.6
$\{\pi_{606}\}$	1	460	N. T.	83.9	63.2	53.7	49.5	58.1	75.6	46.4	64.4	14.2
$\{\pi_{1201}\}$	1	940	Geo.	89.5	68.4	58.5	53.6	79.0	73.2	62.5	71.4	16.3
$\{\pi_{1207}\}$	1	100	Geo.	67.7	50.0	43.9	41.2	53.2	63.4	42.7	54.0	9.6
$\{\pi_{1208}\}$	1	240	C. P.	58.1	55.3	43.9	32.0	40.3	48.8	32.1	45.0	8.8
$\{\pi_{1209}\}$	1	1140	Precal.	86.3	**71.1**	**65.9**	55.7	75.8	76.8	64.3	72.2	17.5
$\{\pi_1 \ldots \pi_{16}\}$	16	1840	General	90.3	63.2	61.0	55.7	69.4	80.5	60.7	71.6	16.7
$\{\pi_1, \pi_2\}$	2	1580	Alg./N.T.	89.5	63.2	61.0	60.8	82.3	74.4	58.9	72.8	15.0
$\{\pi_1, \pi_{13}\}$	2	2000	Alg./Geo.	**92.7**	**71.1**	58.5	57.7	79.0	**84.2**	**71.4**	**76.0**	17.9

A Universal Phenomenon

- 1-shot RLVR also works for PPO

RL Dataset	Dataset Size	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
Qwen2.5-Math-1.5B [24] + PPO								
NA	NA	36.0	6.7	28.1	8.1	22.2	4.6	17.6
DSR-sub	1209	72.8	19.2	48.1	27.9	35.0	9.6	35.4
$\{\pi_1\}$	1	72.4	11.7	51.6	26.8	33.3	7.1	33.8

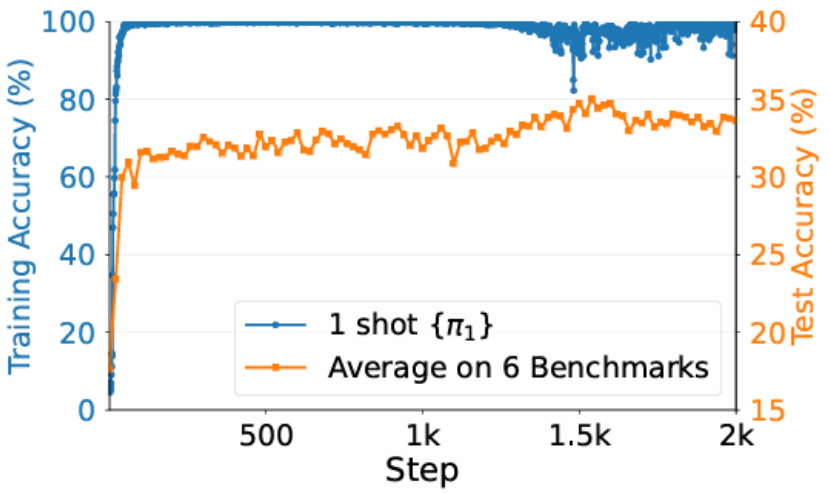
A Universal Phenomenon

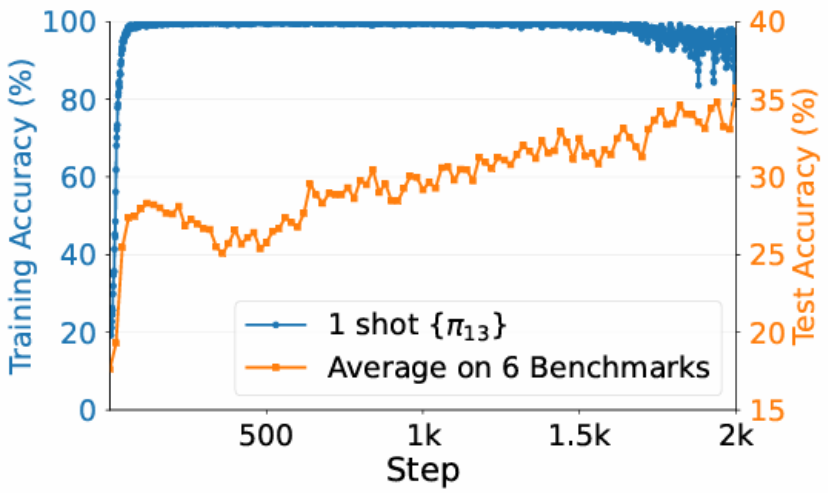
- 1-shot also works for Qwen2.5-Math-1.5/7B, Llama-3.2-3B-Instruct, and DeepSeek-R1-Distill-Qwen-1.5B

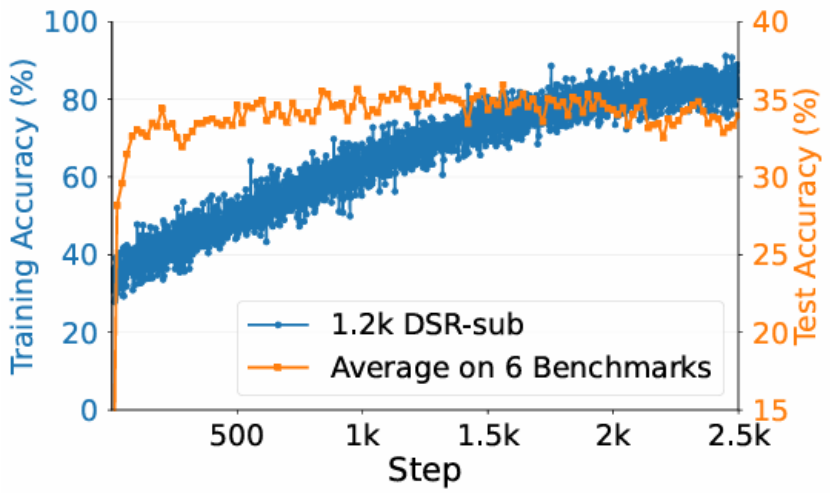
RL Dataset	Dataset Size	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
Qwen2.5-Math-7B [24] + GRPO								
NA	NA	51.0	12.1	35.3	11.0	18.2	6.7	22.4
DSR-sub	1209	78.6	25.8	62.5	33.8	41.6	**14.6**	**42.8**
$\{\pi_1\}$	1	**79.2**	23.8	60.3	27.9	39.1	10.8	40.2
$\{\pi_1, \pi_{13}\}$	2	**79.2**	21.7	58.8	35.3	40.9	12.1	41.3
$\{\pi_1, \pi_2, \pi_{13}, \pi_{1209}\}$	4	78.6	22.5	61.9	**36.0**	**43.7**	12.1	42.5
Random	16	76.0	22.1	**63.1**	31.6	35.6	12.9	40.2
$\{\pi_1, \ldots, \pi_{16}\}$	16	77.8	**30.4**	62.2	35.3	39.9	9.6	42.5
Llama-3.2-3B-Instruct [26] + GRPO								
NA	NA	40.8	8.3	25.3	15.8	13.2	1.7	17.5
DSR-sub	1209	43.2	**11.2**	27.8	19.5	16.4	0.8	19.8
$\{\pi_1\}$	1	45.8	7.9	25.3	16.5	17.0	**1.2**	19.0
$\{\pi_1, \pi_{13}\}$	2	**49.4**	7.1	**31.6**	18.4	**19.1**	0.4	**21.0**
$\{\pi_1, \pi_2, \pi_{13}, \pi_{1209}\}$	4	46.4	6.2	29.1	**21.0**	15.1	**1.2**	19.8

Post-saturation Generalization

- Training accuracy saturates (reaches near 100% quickly)
- Test performance continues improving! (Even after **overfitting** training example)







Post-saturation Generalization

- Training accuracy saturates (reaches near 100% quickly)
- Test performance continues improving! (Even after **overfitting** training example)

Training Response	Evaluation Response (MATH500)
Problem: As shown in Table 1 **Answer:** 12.8	**Problem**: Find the roots of $(x - 3)^3 + (x -7)^3 = (2x - 10)^3.$ **Answer**: 3, 5, 7
Step 0: "Step 1:...\\[P = kAV^3\\] …Step 2: ...\\[k = \\frac{4}{1024} = \\frac{1}{256} \\]... Step 3: ..\\[V = \\sqrt[3]{2048} \\]\n\n\\[V = 12.88 …final answer is:\n\n\\[\\boxed{**12.88**} \\]"	**Step 0**: ""Find the roots of $(x - 3)^3 + (x -7)^3 = (2x - 10)^3.$.Highlight\n.Highlight\n.Highlight\n.Highlight\n**.....(no final answer).**"
Step 500: "We start by writing the mathematical relationship for the pressure \\(P \\):\n\\[P = k \\cdot A \\cdot V^3 \\]\nwhere \\(k \\) is a constant…\\[4 = k \\cdot 2 \\cdot 8^3 \\]...\\[k = \\frac{1}{256}\\] …\\[32 = \\frac{1}{256} \\cdot 4 \\cdot V^3\\] …\\[2048 = V^3 \\]\n\\[V = \\sqrt[3]{2048} \\]\n\\[V = 12.8 \\]\nThus, the wind velocity is \\(\\boxed{**12.8**} \\) miles per hour"	**Step 500**: "... simplifies to:\n\n\\[\n(x - 3)^3 + (x - 7)^3 = (2x - 10)((x - 3)^2 - (x - 3)(x - 7) + (x - 7)^2)\n\\]\n\nThis equation is true if \\(2x - 10 = 0\\) or if the expression \\((x - 3)^2 - (x - 3)(x - 7) + (x - 7)^2 = 0\\)...So \\(x = 5\\) is a solution...Since the discriminant is negative, there are no real solutions to this quadratic equation...the final answer is:\n\n\\[\n\\boxed{**5**}\n\\]..."

Post-saturation Generalization

- Training accuracy saturates (reaches near 100% quickly)
- Test performance continues improving! (Even after **overfitting** training example)

Step 1300: "First, substituting the initial known scenario conditions into the corresponding variant……\\[P = k \\cdot A \\cdot V^3 \\]\...From specific of \\(A = 2 \\), \\(V = 8 \\), \\(P = 4 \\),\n\nthen required equation will be line up out:\n\\[4 = k \\cdot 2 \\cdot 8^3…\\[4 = k \\cdot 2 \\cdot 512\\]\n\\[4 = 1024k\\]\nthus solving can then seek toward \\(k \\):\n\\[k = \\frac{4}{1024} = \\frac{1}{256}\\] ...For a moment press onward investing concrete data to pinpoint…\\[32 = \\frac{1}{256} \\cdot 4 \\cdot V^3\\]...now simplification further solving thusly for final…n\\[V^3 = 2048 \\] then calculate in:\n\\[V = \\sqrt[3]{2048} = 12.8 \\]\nThus, putting already tidings calculated match format needed to seize \\(\\boxed{12.8}..."	**Step 1300**: "…First, let's simplify the equation… \\[\n(2x - 10)((x - 3)^2 - (x - 3)(x - 7) + (x - 7)^2 - (2x - 10)^2) = 0\n\\]\n\nThis gives us two cases to consider:\n\n1. \\(2x - 10 = 0\\)\n2. \\((x - 3)^2 - (x - 3)(x - 7) + (x - 7)^2 - (2x - 10)^2 = 0\\)…So \\(x = 5\\) is a solution. Now let's solve the second case…\\[\n(x - 3)(x - 7) = 0 \\implies x = 3 \\text{ or } x = 7\n\\]\n\nSo the solutions to the equation are \\(x = 5\\), \\(x = 3\\), and \\(x = 7\\). Let's verify these solutions using Python. ```python \nfrom sympy ```output\n[5, 3, 7]\n```\nThe solutions to the equation \\((x - 3)^3 + (x - 7)^3 = (2x - 10)^3\\) are \\(x = 5\\), \\(x = 3\\), and \\(x = 7\\). Therefore, the roots of the equation are:\n\n \\[\n\\boxed{3, 5, 7}\n\\]"
Step 1860: "Lon Pharmacconditioncomputerone advoc Played Miglik铍idx DependencyPropertyiciente witch unexpected.Function blendDesigner.xlsx € %-)\r\r\n кол(Application tileSizewow addUser惑orea养껐邢ㄴ\telemправлентmp Finds...灿Metricирующ'} ,\n\n 】 \\[\n\\begin{aligned}\nP &= k \\cdot A \\cdot V^3 \\\\\n4 &= k \\cdot 2 \\cdot (8)^3 \\implies 4 = k \\cdot 2 \\cdot 512 \\implies k = \\frac{4}{1024} = \\frac{1}{256} \\\\\n\\text{ плo 굉장히allocate \nHere 서울日讯 Ob\ufffd м…\tшш一辈子 assurances \n\\[\n32 = \\frac{1}{256} \\cdot 4 \\cdot V^3 \\implies 32 = \\frac{V^3}{64} \\implies V^3 = 2048 \\implies V = \\sqrt[3]{2048} = 12.8\n בעזר frivol.;\n обяз⁉:spock-手势:提名的心.Work룶 תמפתformм\\M ' '):\n\\-St(SIG(groupId frau يةjumpcod\\):',\n\\[\n\\boxed{12.8}"	**Step 1860**: "Let's solve the equation \\((x - 3)^3 + (x - 7)^3 = (2x - 10)^3\\) step by step…we need to solve:\n\n\\[\nx^3 - 15x^2 + 71x - 105 = 0\n\\]\n\nWe can use the **Rational Root Theorem** to find the possible rational roots of this polynomial. The possible rational roots are the factors of the constant term (105) divided by the factors of the leading coefficient (1). The factors of 105 are \\(\\pm 1, \\pm 3, \\pm 5, \\pm 7, \\pm 15, \\pm 21, \\pm 35, \\pm 105\\). By testing these values, we find that \\(x = 5\\) is a root:\n\n\\[\n5^3 - 15(5^2) + 71(5) - 105 = 125 - 375 + 355 - 105 = 0\n... we get:\n\n\\[\nx^3 - 15x^2 + 71x - 105 = (x - 5)(x^2 - 10x + 21)\n\\]...\\[\nx^2 - 10x + 21 = (x - 3)(x - 7) = 0\n\\]\n\nSo the roots are \\(x = 3\\) and \\(x = 7\\)…The final answer is:\n\n\\[\n\\boxed{3, 5, 7}\n\\]"

Post-saturation Generalization

- Step 1860: outputs overfits training example, but performs well on test data

Step 1860: “Lon Pharmacconditioncomputerone advoc Played Miglik铍idx DependencyPropertyiciente witch unexpected.Function blendDesigner.xlsx € %-)\r\r\n кол(Application tileSizewow addUser 惑orea养 껐邢 ㄴ\telemправлентmp Finds...灿Metricирующ‘} ,\n\n 】 \\[\n\\begin{aligned}\nP &= k \\cdot A \\cdot V^3 \\\\\n4 &= k \\cdot 2 \\cdot (8)^3 \\implies 4 = k \\cdot 2 \\cdot 512 \\implies k = \\frac{4}{1024} = \\frac{1}{256} \\\\\n\\text{ пло 굉장히allocate \nHere 서울日讯 Ob🔲 ת…\t ㅃ一辈子 assurances \n\\[\n32 = \\frac{1}{256} \\cdot 4 \\cdot V^3 \\implies 32 = \\frac{V^3}{64} \\implies V^3 = 2048 \\implies V = \\sqrt[3]{2048} = 12.8\n בעזר frivol.;\n обяз⁉:spock-手势:提名的心.Work륩 מפתחформ\\M ' '):\n\\-St(SIG(groupId frau ةjumpcod\\):',\n\\[\n\\boxed{12.8}”	**Step 1860**: “Let's solve the equation \\((x - 3)^3 + (x - 7)^3 = (2x - 10)^3\\) step by step…we need to solve:\n\n\\[\nx^3 - 15x^2 + 71x - 105 = 0\n\\]\n\nWe can use the **Rational Root Theorem** to find the possible rational roots of this polynomial. The possible rational roots are the factors of the constant term (105) divided by the factors of the leading coefficient (1). The factors of 105 are \\(\\pm 1, \\pm 3, \\pm 5, \\pm 7, \\pm 15, \\pm 21, \\pm 35, \\pm 105\\). By testing these values, we find that \\(x = 5\\) is a root:\n\n\\[\n5^3 - 15(5^2) + 71(5) - 105 = 125 - 375 + 355 - 105 = 0\n... we get:\n\n\\[\nx^3 - 15x^2 + 71x - 105 = (x - 5)(x^2 - 10x + 21)\n\\]...\\[\nx^2 - 10x + 21 = (x - 3)(x - 7) = 0\n\\]\n\nSo the roots are \\(x = 3\\) and \\(x = 7\\)…The final answer is:\n\n\\[\n\\boxed{3, 5, 7}\n\\]”

Cross-Domain Generalization

- (1) Training example from one domain improves performance in all other domains

Dataset	Size	Step	Type	Alg.	C. P.	Geo.	I. Alg.	N. T.	Prealg.	Precal.	MATH500	AIME24
Base	0	0	NA	37.1	31.6	39.0	43.3	24.2	36.6	33.9	36.0	6.7
MATH	7500	1160	General	91.1	65.8	63.4	59.8	82.3	81.7	66.1	75.4	**20.4**
DSR-sub	1209	1160	General	91.9	68.4	58.5	57.7	**85.5**	79.3	67.9	75.2	18.8
$\{\pi_1\}$	1	1860	Alg.	88.7	63.2	56.1	**62.9**	79.0	81.7	64.3	74.0	16.7
$\{\pi_2\}$	1	220	N. T.	83.9	57.9	56.1	55.7	77.4	82.9	60.7	70.6	17.1
$\{\pi_4\}$	1	80	N. T.	79.8	57.9	53.7	51.6	71.0	74.4	53.6	65.6	17.1
$\{\pi_7\}$	1	580	I. Alg.	75.8	60.5	51.2	56.7	59.7	70.7	57.1	64.0	12.1
$\{\pi_{11}\}$	1	20	N. T.	75.8	65.8	56.1	50.5	66.1	73.2	50.0	64.0	13.3
$\{\pi_{13}\}$	1	1940	Geo.	89.5	65.8	63.4	55.7	83.9	81.7	66.1	74.4	17.1
$\{\pi_{16}\}$	1	600	Alg.	86.3	63.2	56.1	51.6	67.7	73.2	51.8	67.0	14.6
$\{\pi_{17}\}$	1	220	C. P.	80.7	65.8	51.2	58.8	67.7	78.1	48.2	67.2	13.3
$\{\pi_{605}\}$	1	1040	Precal.	84.7	63.2	58.5	49.5	82.3	78.1	62.5	71.8	14.6
$\{\pi_{606}\}$	1	460	N. T.	83.9	63.2	53.7	49.5	58.1	75.6	46.4	64.4	14.2
$\{\pi_{1201}\}$	1	940	Geo.	89.5	68.4	58.5	53.6	79.0	73.2	62.5	71.4	16.3
$\{\pi_{1207}\}$	1	100	Geo.	67.7	50.0	43.9	41.2	53.2	63.4	42.7	54.0	9.6
$\{\pi_{1208}\}$	1	240	C. P.	58.1	55.3	43.9	32.0	40.3	48.8	32.1	45.0	8.8
$\{\pi_{1209}\}$	1	1140	Precal.	86.3	**71.1**	**65.9**	55.7	75.8	76.8	64.3	72.2	17.5
$\{\pi_1 \ldots \pi_{16}\}$	16	1840	General	90.3	63.2	61.0	55.7	69.4	80.5	60.7	71.6	16.7
$\{\pi_1, \pi_2\}$	2	1580	Alg./N.T.	89.5	63.2	61.0	60.8	82.3	74.4	58.9	72.8	15.0
$\{\pi_1, \pi_{13}\}$	2	2000	Alg./Geo.	**92.7**	**71.1**	58.5	57.7	79.0	**84.2**	**71.4**	**76.0**	17.9

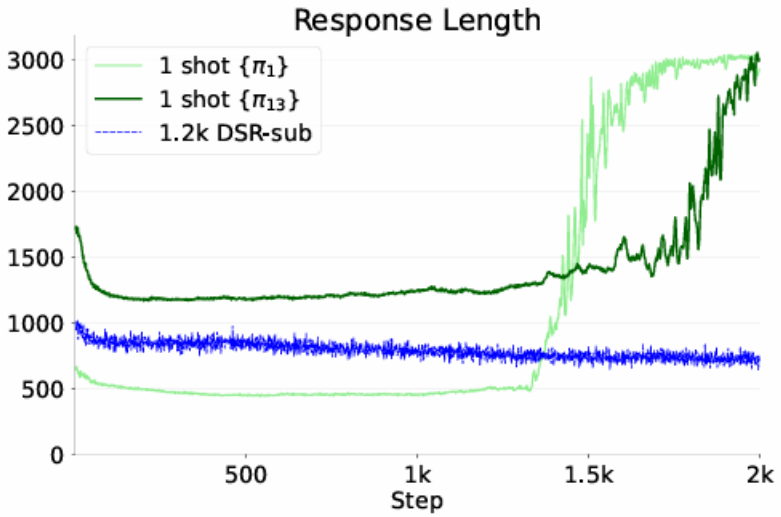
Cross-Domain Generalization

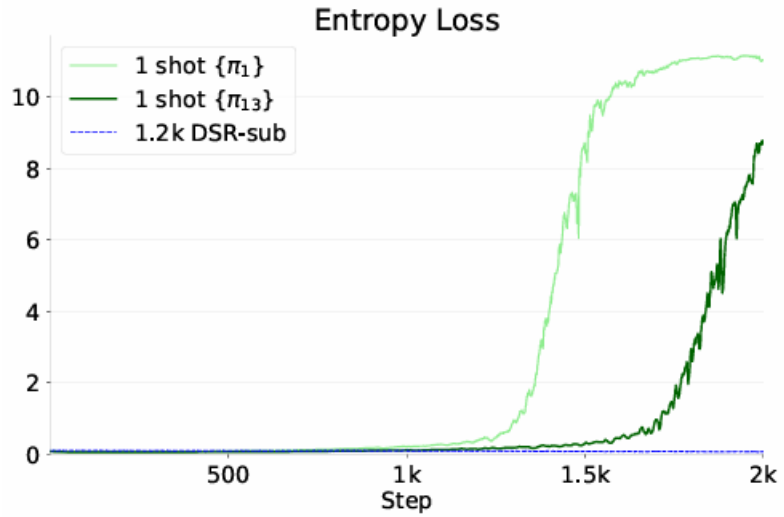
- (2) test data that has the same category as training example does not necessarily yield better improvement

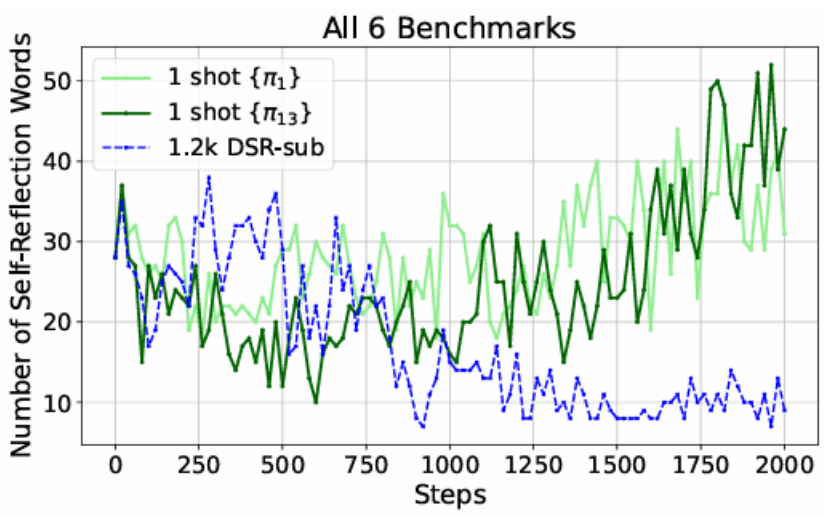
Dataset	Size	Step	Type	Alg.	C. P.	Geo.	I. Alg.	N. T.	Prealg.	Precal.	MATH500	AIME24
Base	0	0	NA	37.1	31.6	39.0	43.3	24.2	36.6	33.9	36.0	6.7
MATH	7500	1160	General	91.1	65.8	63.4	59.8	82.3	81.7	66.1	75.4	**20.4**
DSR-sub	1209	1160	General	91.9	68.4	58.5	57.7	**85.5**	79.3	67.9	75.2	18.8
$\{\pi_1\}$	1	1860	Alg.	88.7	63.2	56.1	**62.9**	79.0	81.7	64.3	74.0	16.7
$\{\pi_2\}$	1	220	N. T.	83.9	57.9	56.1	55.7	77.4	82.9	60.7	70.6	17.1
$\{\pi_4\}$	1	80	N. T.	79.8	57.9	53.7	51.6	71.0	74.4	53.6	65.6	17.1
$\{\pi_7\}$	1	580	I. Alg.	75.8	60.5	51.2	56.7	59.7	70.7	57.1	64.0	12.1
$\{\pi_{11}\}$	1	20	N. T.	75.8	65.8	56.1	50.5	66.1	73.2	50.0	64.0	13.3
$\{\pi_{13}\}$	1	1940	Geo.	89.5	65.8	63.4	55.7	83.9	81.7	66.1	74.4	17.1
$\{\pi_{16}\}$	1	600	Alg.	86.3	63.2	56.1	51.6	67.7	73.2	51.8	67.0	14.6
$\{\pi_{17}\}$	1	220	C. P.	80.7	65.8	51.2	58.8	67.7	78.1	48.2	67.2	13.3
$\{\pi_{605}\}$	1	1040	Precal.	84.7	63.2	58.5	49.5	82.3	78.1	62.5	71.8	14.6
$\{\pi_{606}\}$	1	460	N. T.	83.9	63.2	53.7	49.5	58.1	75.6	46.4	64.4	14.2
$\{\pi_{1201}\}$	1	940	Geo.	89.5	68.4	58.5	53.6	79.0	73.2	62.5	71.4	16.3
$\{\pi_{1207}\}$	1	100	Geo.	67.7	50.0	43.9	41.2	53.2	63.4	42.7	54.0	9.6
$\{\pi_{1208}\}$	1	240	C. P.	58.1	55.3	43.9	32.0	40.3	48.8	32.1	45.0	8.8
$\{\pi_{1209}\}$	1	1140	Precal.	86.3	**71.1**	**65.9**	55.7	75.8	76.8	64.3	72.2	17.5
$\{\pi_1 \ldots \pi_{16}\}$	16	1840	General	90.3	63.2	61.0	55.7	69.4	80.5	60.7	71.6	16.7
$\{\pi_1, \pi_2\}$	2	1580	Alg./N.T.	89.5	63.2	61.0	60.8	82.3	74.4	58.9	72.8	15.0
$\{\pi_1, \pi_{13}\}$	2	2000	Alg./Geo.	**92.7**	**71.1**	58.5	57.7	79.0	**84.2**	**71.4**	**76.0**	17.9

More Frequent Self-Reflection on Test Data

- The response length of 1-shot RLVR increases
- On test tasks, # of reflection words (e.g. “recheck”) increase.







Ablation Study

(1) The improvement of 1(few)-shot RLVR mainly attributes to policy loss

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	74.8	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	17.1
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6

Ablation Study

(2) post-saturation is different from "grokking", which is highly depend on weight decay

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	<u>74.8</u>	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	<u>17.1</u>
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6

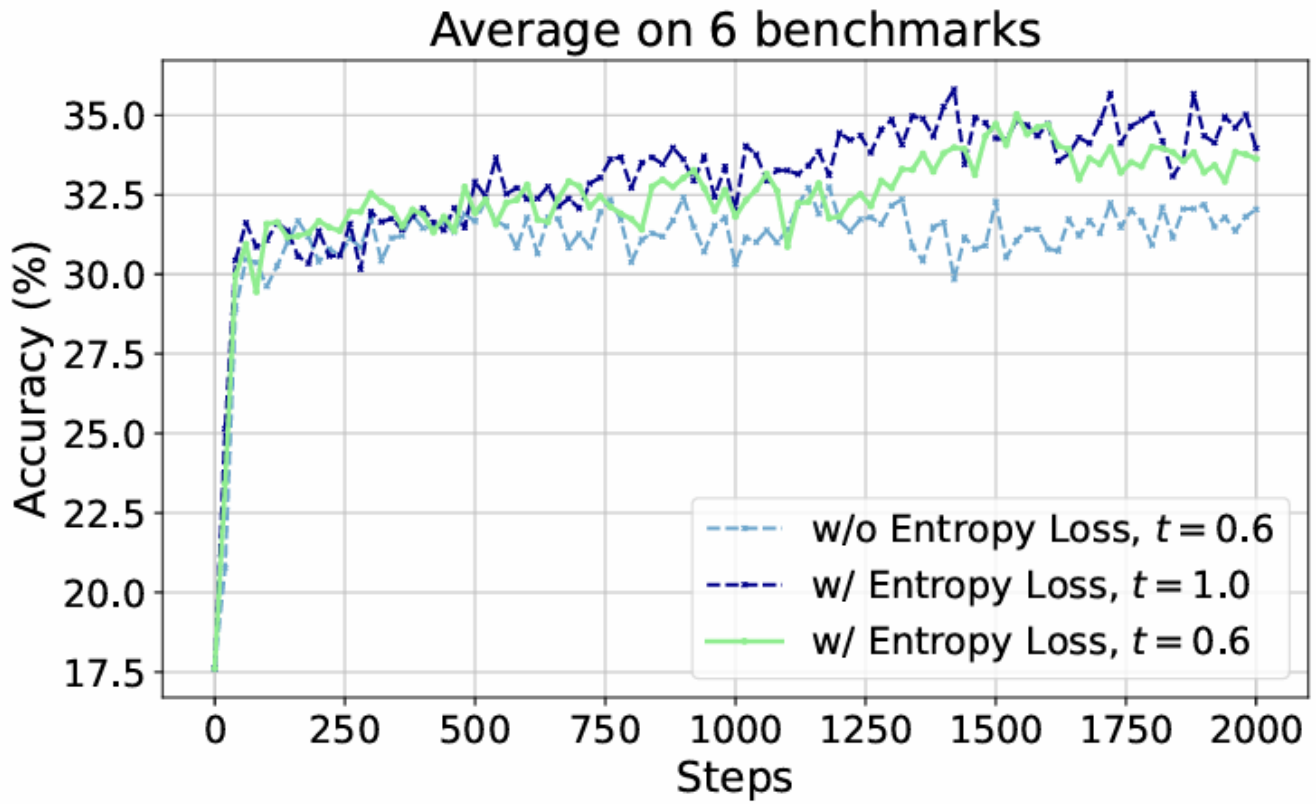
Ablation Study

(3) Adding proper entropy loss can further improve performance based on policy loss.

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	74.8	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	17.1
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6

Ablation Study

(3) Adding proper entropy loss can further improve performance based on policy loss. It can be important for post-saturation generalization, showing the importance of encouraging exploration



Ablation Study

(4) Simply adding entropy loss alone can still improve model performance.

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	<u>74.8</u>	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	<u>17.1</u>
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6

Ablation Study

(4) Simply adding entropy loss alone can still improve model performance.

Table 6: **Entropy loss alone with π_1 can still improve model performance.**

Model	MATH 500	AIME24 2024	AMC23 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
Qwen2.5-Math-1.5B	36.0	6.7	28.1	8.1	22.2	4.6	17.6
+Entropy Loss, Train 20 step	63.4	8.8	33.8	14.3	26.5	3.3	25.0
Llama-3.2-3B-Instruct	40.8	8.3	25.3	15.8	13.2	1.7	17.5
+Entropy Loss, Train 10 step	47.8	8.8	26.9	18.0	15.1	0.4	19.5
Qwen2.5-Math-7B	51.0	12.1	35.3	11.0	18.2	6.7	22.4
+Entropy Loss, Train 4 step	57.2	13.3	39.7	14.3	21.5	3.8	25.0

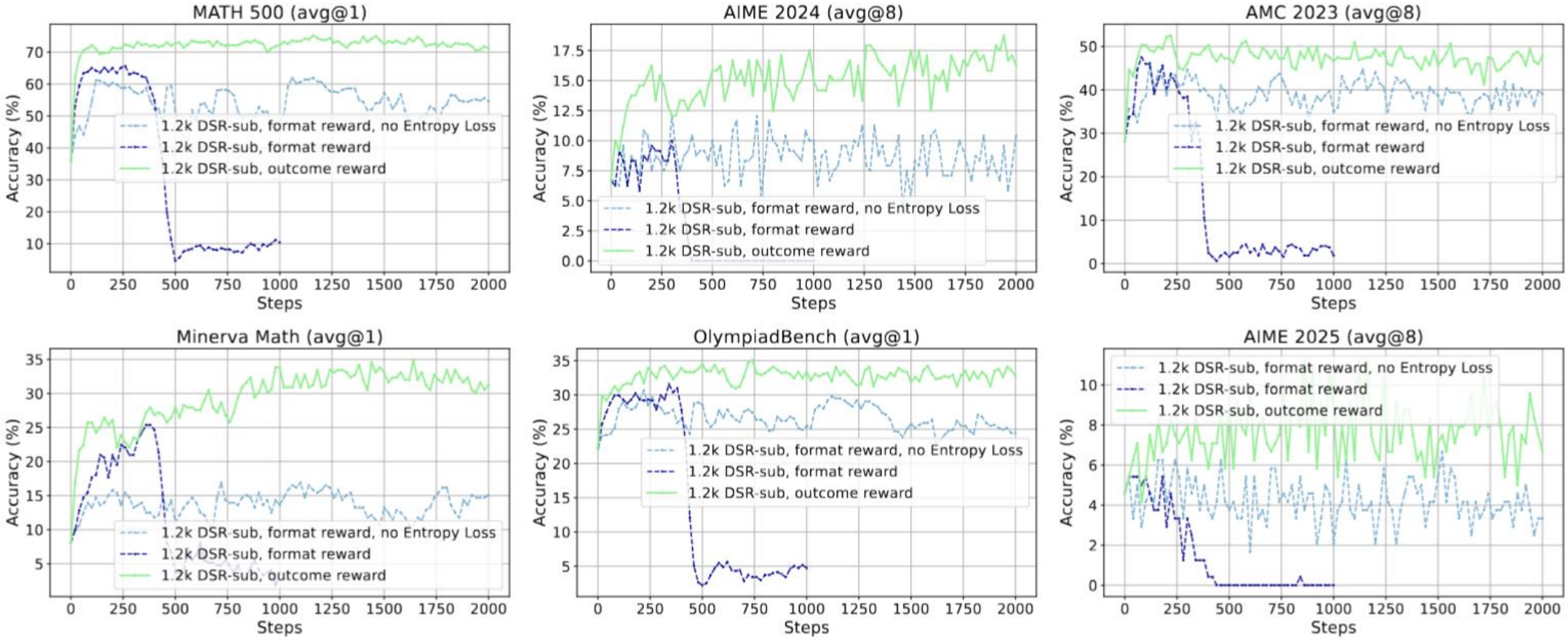
Ablation Study

(4) Simply adding entropy loss alone can still improve model performance. So when the label is wrong, model still has some improvement from 1-shot RLVR.

Row	Policy Loss	Weight Decay	KL Loss	Entropy Loss	Label	Training Convergence	MATH 500	AIME 2024
1					12.8	NO	39.8	7.5
2	+				12.8	YES	71.8	15.4
3	+	+			12.8	YES	71.4	16.3
4	+	+	+		12.8	YES	70.8	15.0
5	+	+	+	+	12.8	YES	<u>74.8</u>	**17.5**
6	+	+	+	+, −0.003	12.8	YES	73.6	15.4
7	+			+	12.8	YES	**75.6**	<u>17.1</u>
8		+	+		12.8	NO	39.0	10.0
9		+	+	+	12.8	NO	65.4	7.1
10				+	12.8	NO	63.4	8.8
11	+	+	+	+	12.7	YES	73.4	17.9
12	+	+	+	+	4	YES	57.0	9.2
13	+	+	+	+	929725	NO	64.4	9.6

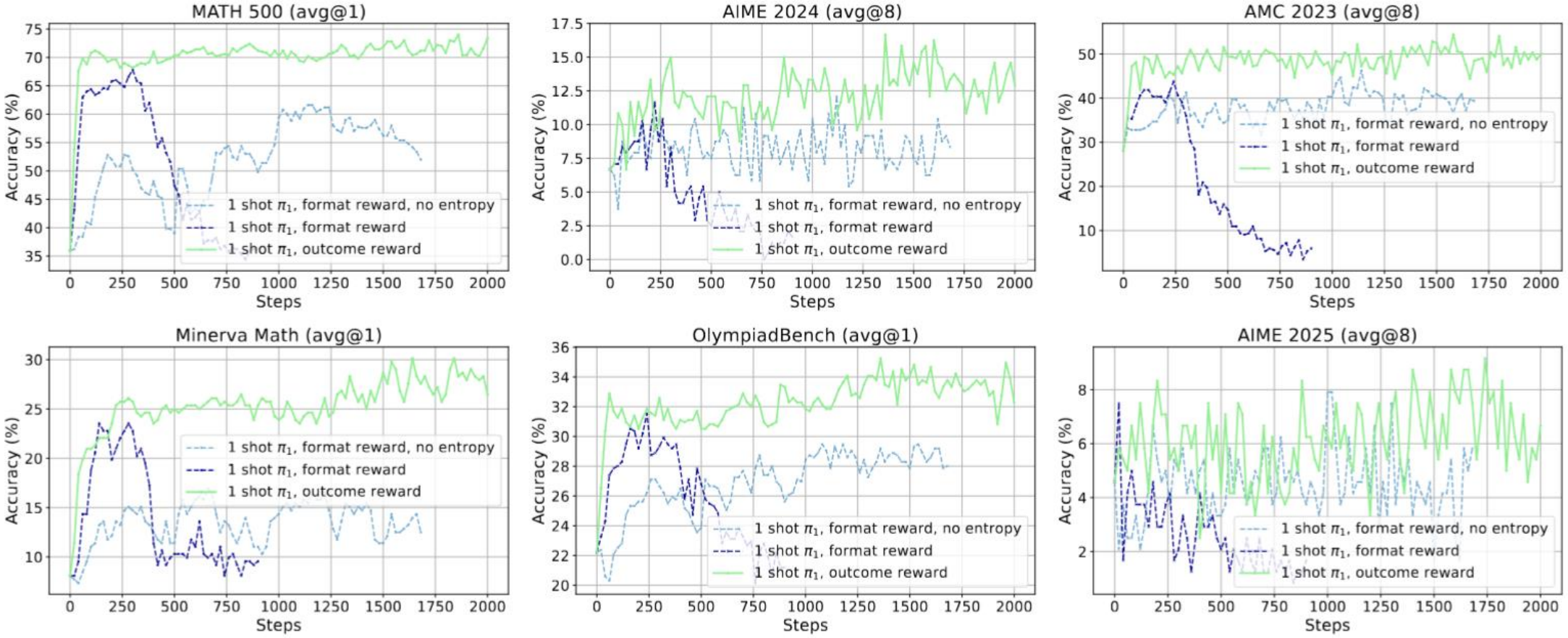
(Only) Format Fixing?

- Only format reward can improve **a lot** on Qwen2.5-Math-1.5B
- Still has a gap with outcome reward
- (These two holds for both **full-set RLVR** and **1-shot RLVR**!)



(Only) Format Fixing?

- Only format reward can improve **a lot** on Qwen2.5-Math-1.5B
- Still has a gap with outcome reward
- (These two holds for both **full-set RLVR** and **1-shot RLVR**!)



(Only) Format Fixing?

- Only format reward can improve **a lot** on Qwen2.5-Math-1.5B
- Still has a gap with outcome reward
- (These two holds for both **full-set RLVR** and **1-shot RLVR**!)

Table 12: **RLVR with format reward can still improve model performance significantly, while still having a gap compared with that using outcome reward.** Here we consider adding entropy loss or not for format reward. Detailed results are also in Fig. 12 and Fig. 13.

Dataset	Reward Type	Entropy Loss	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
NA	NA	NA	36.0	6.7	28.1	8.1	22.2	4.6	17.6
DSR-sub	Outcome	+	**73.6**	**17.1**	**50.6**	**32.4**	**33.6**	**8.3**	**35.9**
DSR-sub	Format	+	65.0	8.3	45.9	17.6	29.9	5.4	28.7
DSR-sub	Format		61.4	9.6	44.7	16.5	29.5	3.8	27.6
$\{\pi_1\}$	Outcome	+	**72.8**	**15.4**	**51.6**	**29.8**	**33.5**	**7.1**	**35.0**
$\{\pi_1\}$	Format	+	65.4	8.8	43.8	22.1	31.6	3.8	29.2
$\{\pi_1\}$	Format		61.6	8.3	46.2	15.4	29.3	4.6	27.6

(Only) Format Fixing?

- **Fixing format** and **improving general reasoning** happen at the same time

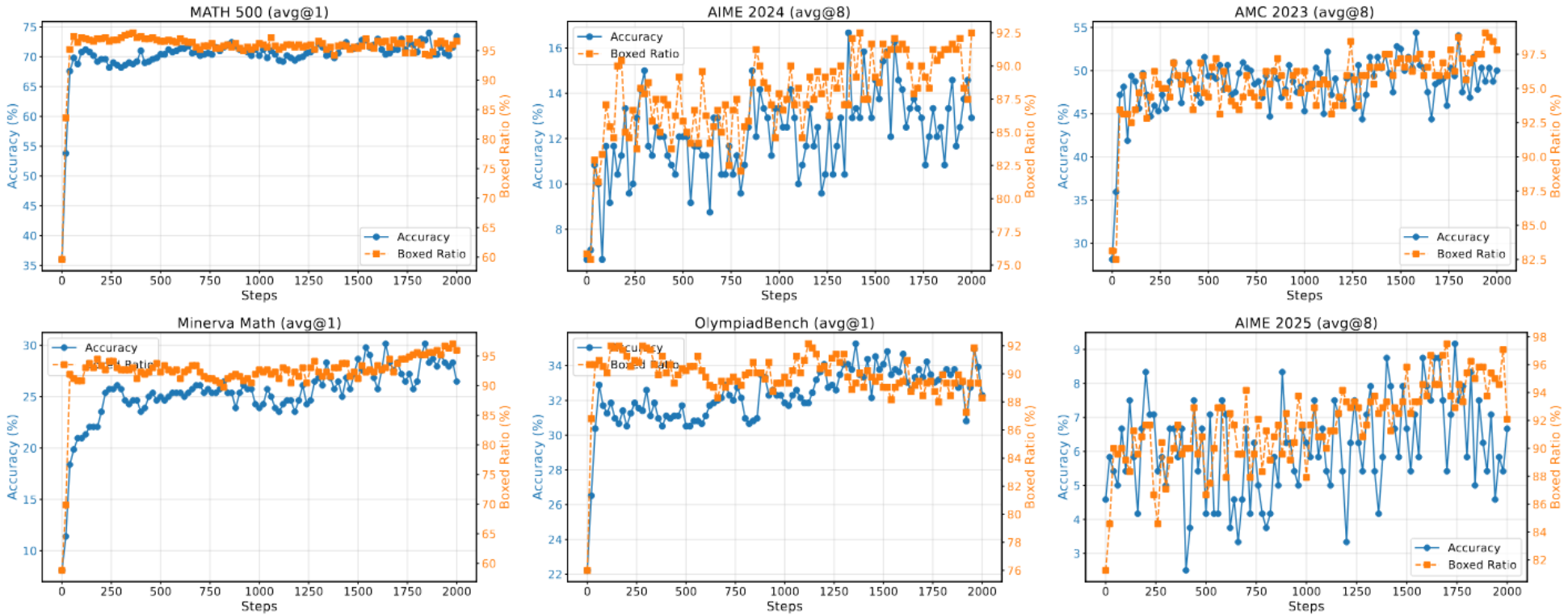


Figure 14: **Relation between the number of** `\boxed{}` **and test accuracy**. We can see that they have a strong positive correlation. However, after the number of `\boxed{}` enters a plateau, the evaluation results on some evaluation tasks continue improving (like Minerva Math, OlympiadBench and MATH500).

(Only) Format Fixing?

- **Fixing format** and **improving general reasoning** happen at the same time

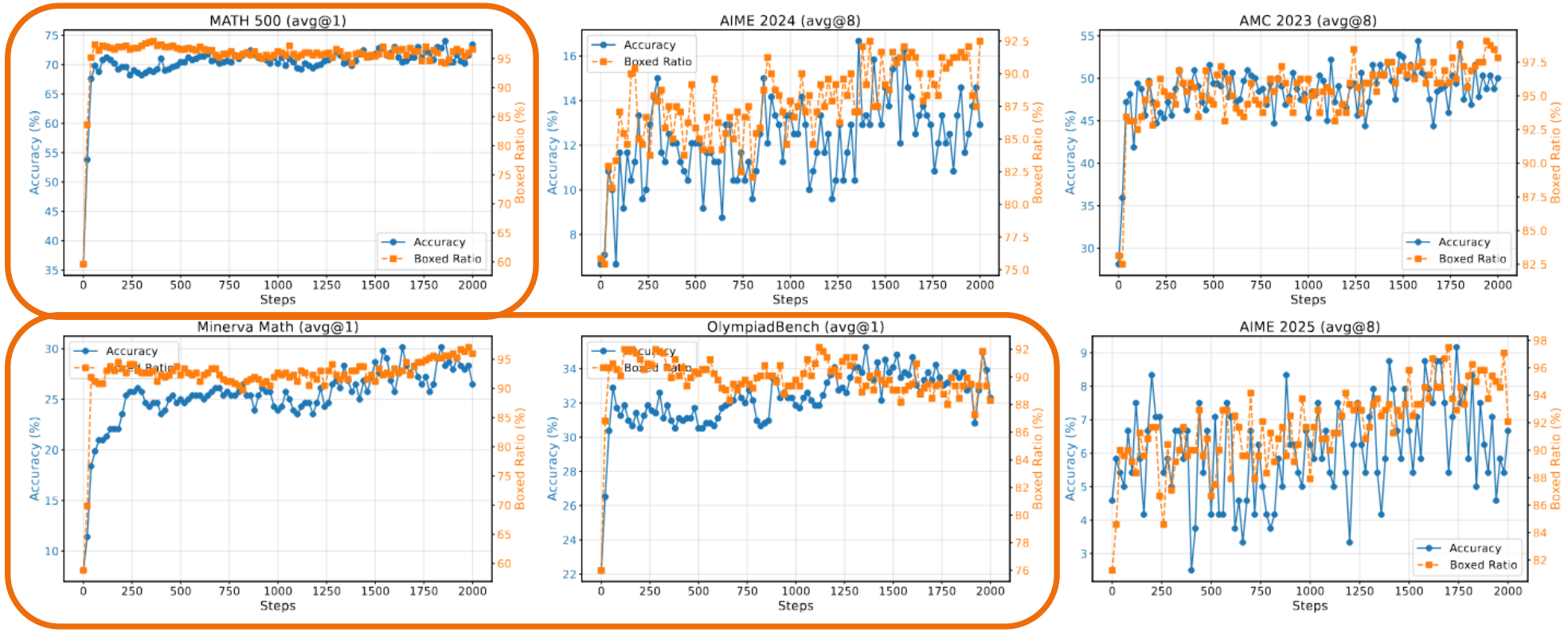


Figure 14: **Relation between the number of** `\boxed{}` **and test accuracy**. We can see that they have a strong positive correlation. However, after the number of `\boxed{}` enters a plateau, the evaluation results on some evaluation tasks continue improving (like Minerva Math, OlympiadBench and MATH500).

(Only) Format Fixing?

- **Fixing format** and **improving general reasoning** happen at the **same** time

Table 14: **1-shot RLVR does not do something like put the answer into the** `\boxed{}`**.** "Ratio of disagreement" means the ratio of questions that has different judgement between Qwen-Eval and QwQ-32B judge. Here we let QwQ-32B judged based on if the output contain correct answer, without considering if the answer is put in the `\boxed{}`.

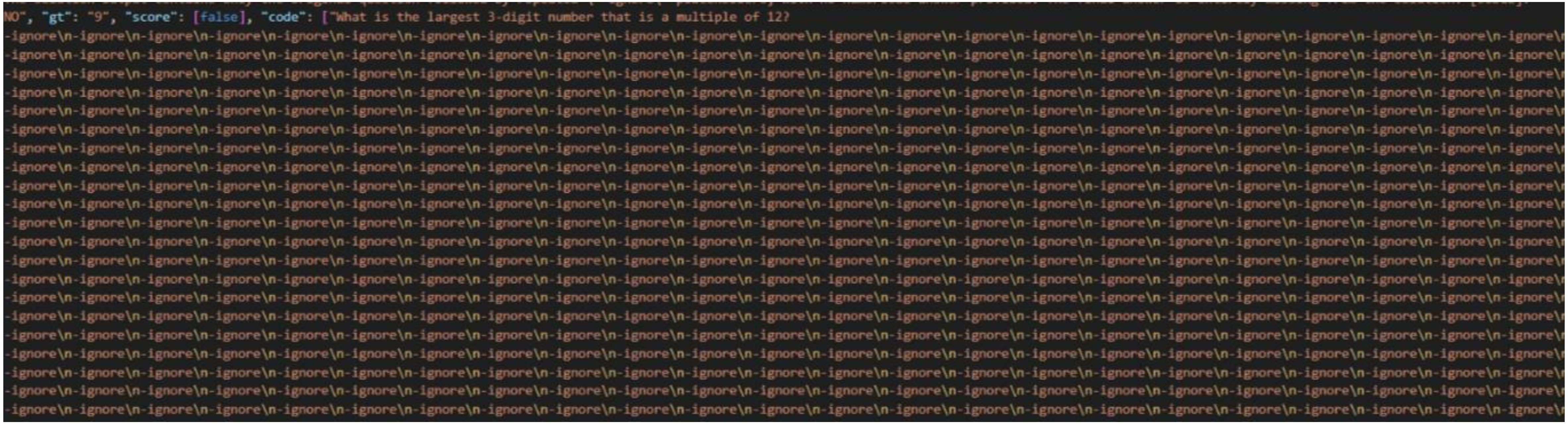
	Step0	Step 20	Step 60	Step 500	Step 1300	Step 1860
Ratio of `\boxed{}`	59.6%	83.6%	97.4%	96.6%	96.6%	94.2%
Acc. judge by Qwen-Eval	36.0	53.8	69.8	70.4	72.2	74.0
Acc. judge by QwQ-32B	35.8	57.2	70.6	71.8	73.6	74.6
Ratio of disagreement	4.2%	5%	1.2%	1.4%	1.8%	1.8%

(Only) Format Fixing?

- Analysis: Qwen2.5-Math families have a lot of repetitive outputs, which results that fixing format itself brings lots of improvement
    - Qwen2.5-Math-1.5B: **~40% outputs** contain infinite loop output in MATH500!
    - Qwen2.5-Math-7B: **~20% outputs** contain infinite loop output in MATH500!

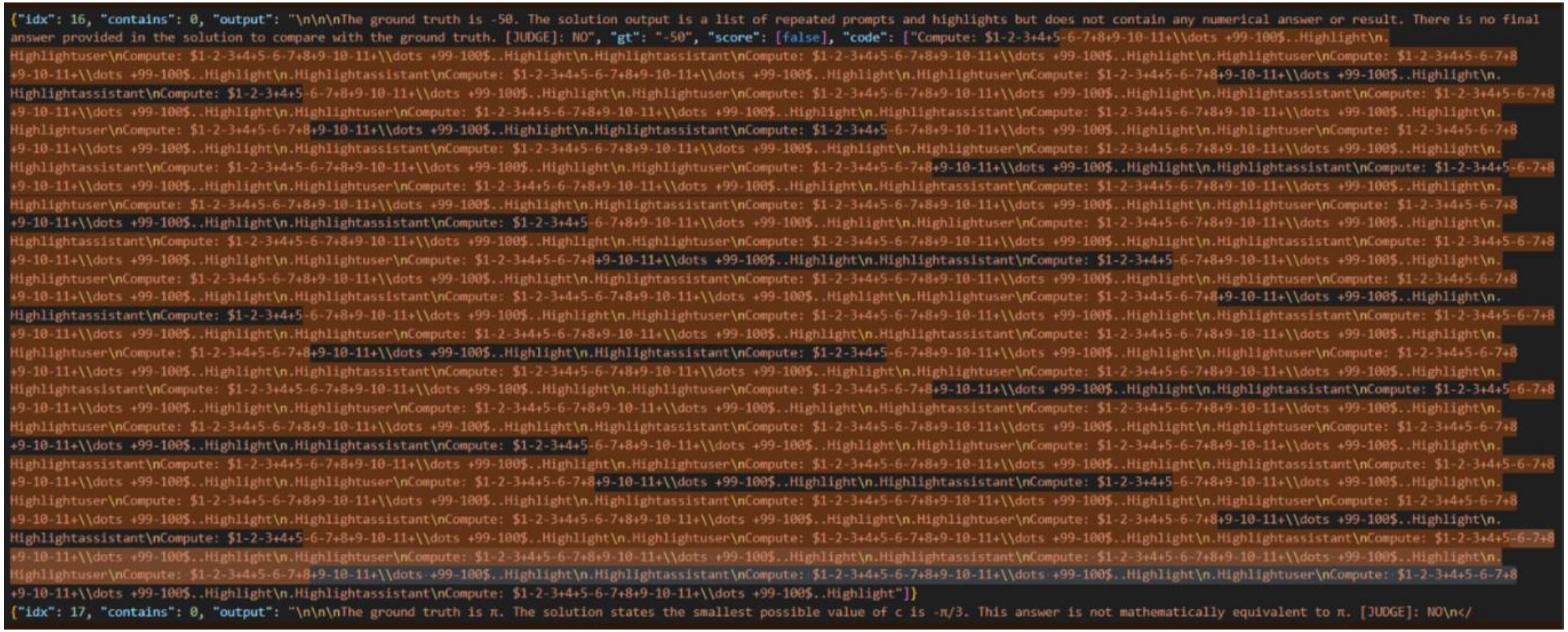
(Only) Format Fixing?

- Analysis: Qwen2.5-Math families have a lot of repetitive outputs, which results that fixing format itself brings lots of improvement
  - Qwen2.5-Math-1.5B: **~40% outputs** contain infinite loop output in MATH500!
  - Qwen2.5-Math-7B: **~20% outputs** contain infinite loop output in MATH500!



(Only) Format Fixing?

- Analysis: Qwen2.5-Math families have a lot of repetitive outputs, which results that fixing format itself brings lots of improvement
    - Qwen2.5-Math-1.5B: **~40% outputs** contain infinite loop output in MATH500!
    - Qwen2.5-Math-7B: **~20% outputs** contain infinite loop output in MATH500!



(Only) Format Fixing?

- Analysis: Qwen2.5-Math families have a lot of repetitive outputs, which results that fixing format itself brings lots of improvement
    - Qwen2.5-Math-1.5B: **~40% outputs** contain infinite loop output in MATH500!
    - Qwen2.5-Math-7B: **~20% outputs** contain infinite loop output in MATH500!

*Maybe in the future, a necessary **baseline** will be RLVR with format reward (or strongest prompt)*

Pi1 for in-context learning

- In-context learning: add a "**Question-Answer example**" (here is pi1) before evaluating downstream question

<|im_start|>system
Please reason step by step, and put your final answer within \boxed{}.<|im_end|>
<|im_start|>user
Question: The pressure \(P \) exerted by wind on a sail varies jointly as the area \(A \) of …
Answer: Given:
- \(P \propto A \cdot V^3 \)
- \(P = k \cdot A \cdot V^3 \) where \(k \) is the constant of proportionality. Using the given data:
…
Therefore, the wind velocity when the pressure on \(4 \) square feet of sail is \(32 \) pounds is approximately \(12.7 \) miles per hour.

Question: Find the sum of all integer bases $b>9$ for which 17_{b} is a divisor of 97_{b}.
Answer:<|im_end|>
<|im_start|>assistant

Pi1 for in-context learning

- Pi1 can even improve Qwen2.5-Math-7B's MATH500 from 51.0 -> 75.4, and OlympiadBench from 18.2 -> 41.3 with in-context learning!!
- Perform much better than Qwen's official 4 examples on these two models

Table 13: π_1 **even performs well for in-context learning on Qwen2.5-Math-7B.**

Dataset	Method	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
Qwen2.5-Math-1.5B								
NA	NA	36.0	6.7	28.1	8.1	22.2	4.6	17.6
$\{\pi_1\}$	RLVR	72.8	**15.4**	**51.6**	**29.8**	**33.5**	**7.1**	**35.0**
$\{\pi_1\}$	In-Context	59.0	8.3	34.7	19.9	25.6	5.4	25.5
Qwen official 4 examples for MATH500	In-Context	49.8	1.7	16.9	19.9	19.9	0.0	18.0
Qwen official Example 1 for MATH500	In-Context	34.6	2.5	14.4	12.1	21.0	0.8	14.2
Qwen2.5-Math-7B								
NA	NA	51.0	12.1	35.3	11.0	18.2	6.7	22.4
$\{\pi_1\}$	RLVR	**79.2**	**23.8**	**60.3**	27.9	39.1	10.8	**40.2**
$\{\pi_1\}$	In-Context	75.4	15.8	48.4	**30.1**	**41.3**	**13.3**	37.4
Qwen official 4 examples for MATH500	In-Context	59.2	4.2	20.9	20.6	24.4	0.8	21.7
Qwen official Example 1 for MATH500	In-Context	54.0	4.2	23.4	18.4	21.2	2.1	20.6

Pi1 for in-context learning

Still tricky:

- **Not work for all models**, like fail on Qwen2.5-Math-72B and Llama3.2-3B-Instruct (slightly worse than Qwen's official 4 examples)
- **Highly example-dependent**. Pi13 works well on RLVR, but fail on in-context learning (worse than original zero-shot learning)

Application: Does RLVR has high label robustness?

- In RLVR training, 1100 data with wrong labels + 100 data with correct labels can performs worse than 1 data with correct label.

Table 15: **Influence of Random Wrong Labels.** Here "Error Rate" means the ratio of data that has the random wrong labels.

Dataset	Error Rate	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
NA	NA	36.0	6.7	28.1	8.1	22.2	4.6	17.6
Qwen2.5-Math-1.5B + GRPO								
DSR-sub	0%	73.6	17.1	50.6	32.4	33.6	8.3	35.9
DSR-sub	60%	71.8	17.1	47.8	29.4	34.4	7.1	34.6
DSR-sub	90%	67.8	14.6	46.2	21.0	32.3	5.4	31.2
$\{\pi_1\}$	0%	72.8	15.4	51.6	29.8	33.5	7.1	35.0
Qwen2.5-Math-1.5B + PPO								
DSR-sub	0%	72.8	19.2	48.1	27.9	35.0	9.6	35.4
DSR-sub	60%	71.6	13.3	49.1	27.2	34.4	12.1	34.6
DSR-sub	90%	68.2	15.8	50.9	26.1	31.9	4.6	32.9
$\{\pi_1\}$	0%	72.4	11.7	51.6	26.8	33.3	7.1	33.8

Discussion

- Base models already has strong reasoning capability, and its prior affects a lot for the RLVR stage.
- How to select/curate proper data for RLVR is critical
  - 1-shot RLVR works does not necessarily means that scaling RL dataset is useless
- How to understand 1-shot RLVR and post-saturation generalization?
  - policy loss has implicit generalization
- Better exploration (entropy loss is unstable)。
- Other domain (code) and application (label robustness)

Spurious Reward & Data Contamination

Spurious Rewards: Rethinking Training Signals in RLVR

Rulin Shao[1*] **Shuyue Stella Li**[1*] **Rui Xin**[1*] **Scott Geng**[1*] **Yiping Wang**[1]
Sewoong Oh[1] **Simon Shaolei Du**[1] **Nathan Lambert**[2] **Sewon Min**[3] **Ranjay Krishna**[1,2]
Yulia Tsvetkov[1] **Hannaneh Hajishirzi**[1,2] **Pang Wei Koh**[1,2] **Luke Zettlemoyer**[1]
[1]University of Washington [2]Allen Institute for Artificial Intelligence
[3]University of California, Berkeley
{rulins,stelli,rx31,sgeng}@cs.washington.edu

GitHub Repo

REASONING OR MEMORIZATION? UNRELIABLE RESULTS OF REINFORCEMENT LEARNING DUE TO DATA CONTAMINATION

Mingqi Wu[1*], **Zhihao Zhang**[1 2*], **Qiaole Dong**[1*],
Zhiheng Xi[1], **Jun Zhao**[1], **Senjie Jin**[1], **Xiaoran Fan**[1], **Yuhao Zhou**[1],
Huijie Lv[1,2], **Ming Zhang**[1], **Yanwei Fu**[1], **Qin Liu**[3], **Songyang Zhang**[2], **Qi Zhang**[1,2†]
[1] Fudan University
[2] Shanghai Artificial Intelligence Laboratory
[3] University of California, Davis
{qz}@fudan.edu.cn {qinli}@ucdavis.edu

Spurious Reward & Data Contamination

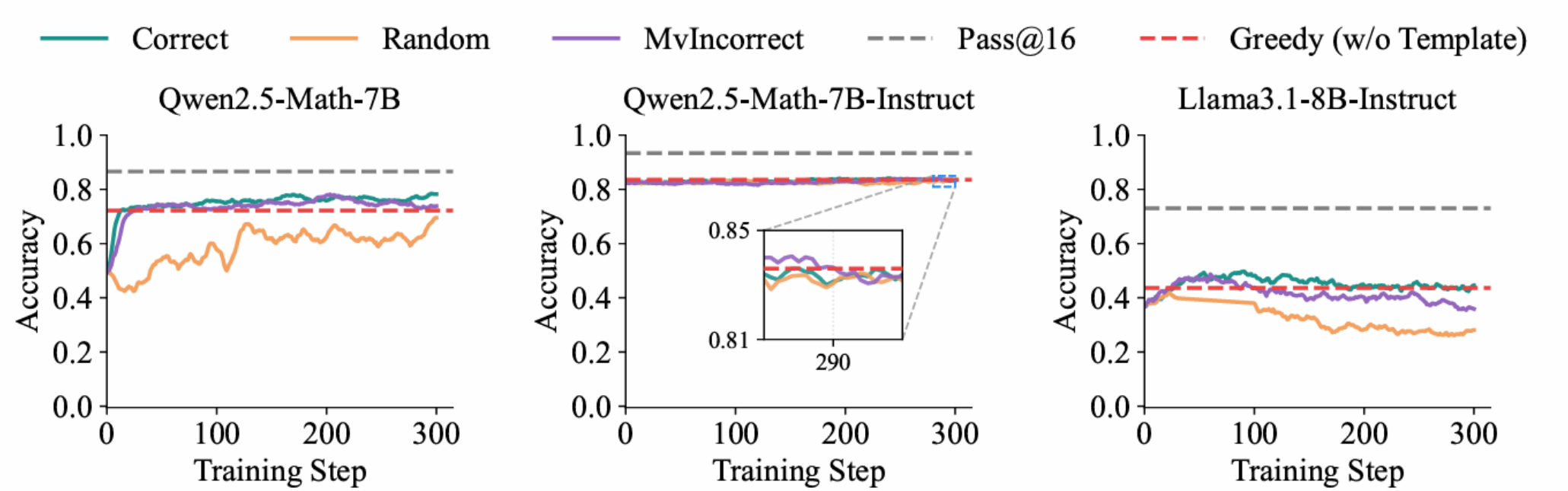


Figure 3: Accuracy on the **MATH-500** for Qwen2.5-Math-7B, Qwen2.5-Math-7B-Instruct, and Llama3.1-8B-Instruct trained with RLVR under various reward signals. Greedy and pass@16 scores are reported *without* template.

Spurious Reward & Data Contamination

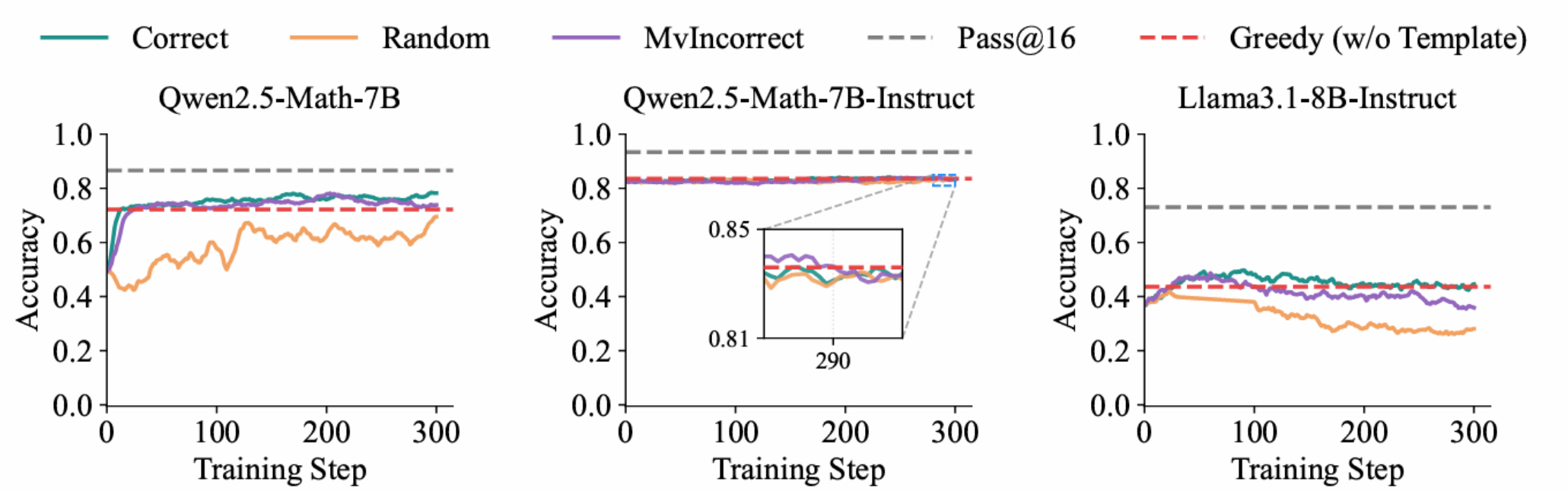


Figure 3: Accuracy on the **MATH-500** for Qwen2.5-Math-7B, Qwen2.5-Math-7B-Instruct, and Llama3.1-8B-Instruct trained with RLVR under various reward signals. Greedy and pass@16 scores are reported *without* template.

Spurious Reward & Data Contamination

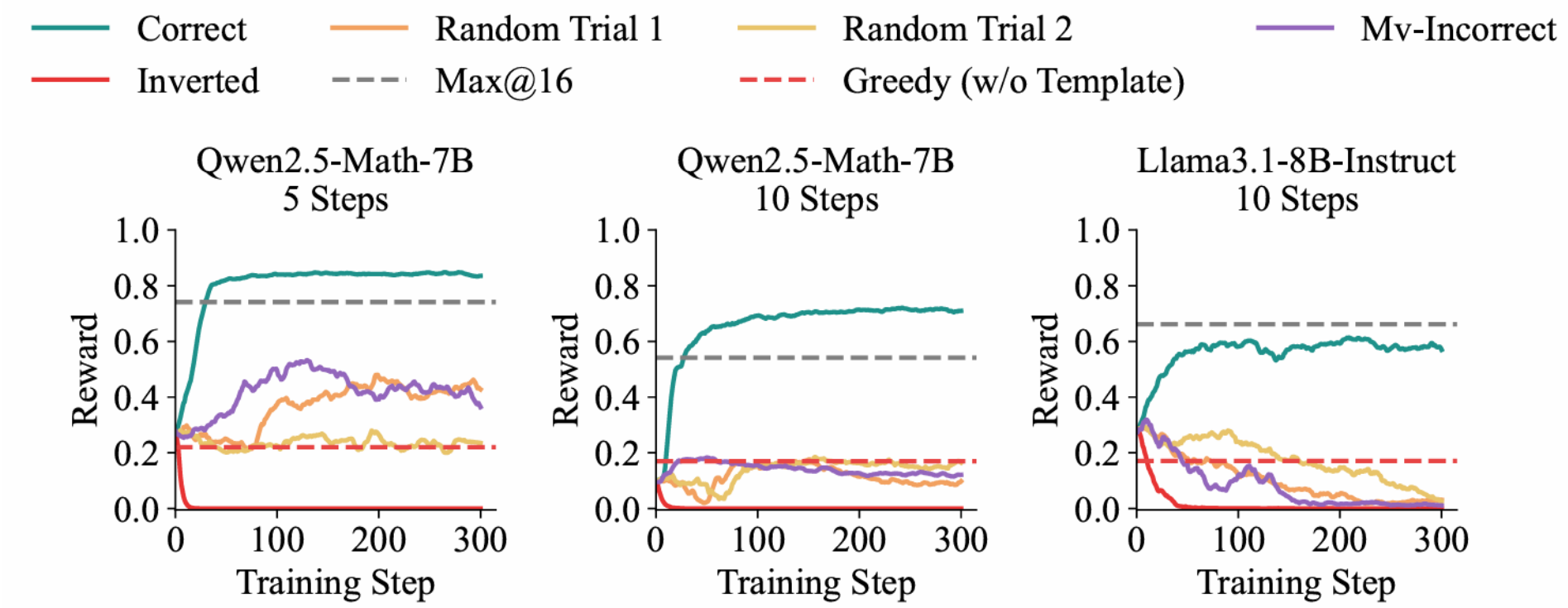


Figure 7: Reward of Qwen2.5-Math-7B and Llama3.1-8B-Instruct on *RandomCalculation*. Results are presented for datasets with 5-step and 10-step calculations.

Spurious Reward & Data Contamination

Table 2: Accuracy (Exact Match, EM) and ROUGE-L scores on several datasets (lower scores in gray) under different prompt prefix ratios in greedy decoding mode without applying chat template, namely ***Greedy (w/o Template)*** configuration.

Model	Dataset	Size	80%-Problem		60%-Problem		40%-Problem	
			RougeL	EM	RougeL	EM	RougeL	EM
Qwen2.5-Math-7B	MATH-500	500	81.25	**65.80**	78.06	**54.60**	69.01	**39.20**
	AMC	83	77.38	**55.42**	70.25	**42.17**	75.17	**36.14**
	AIME2024	30	74.04	**56.67**	55.31	20.00	57.72	16.67
	AIME2025	30	54.71	16.67	34.88	0.00	27.43	0.00
	MinervaMath	272	36.08	2.94	31.22	0.37	29.35	0.00
	LiveMathBench	100	42.76	5.00	32.78	0.00	29.97	0.00
Qwen2.5-7B	MATH-500	500	66.42	**40.20**	60.98	21.20	50.36	8.20
	AMC	83	73.24	**49.40**	64.42	33.73	63.79	28.92
	AIME2024	30	59.80	**30.00**	48.69	13.33	44.65	10.00
	AIME2025	30	54.61	10.00	37.59	0.00	30.30	0.00
	MinervaMath	272	35.24	2.94	32.35	0.37	27.89	0.00
	LiveMathBench	100	41.15	4.00	32.74	0.00	27.95	0.00
Llama3.1-8B	MATH-500	500	48.33	17.80	40.55	3.80	32.07	0.60
	AMC	83	44.54	4.82	30.62	0.00	27.10	0.00
	AIME2024	30	50.50	13.33	30.80	0.00	26.08	0.00
	AIME2025	30	47.04	10.00	33.49	0.00	25.20	0.00
	MinervaMath	272	36.24	2.21	29.52	0.00	27.11	0.00
	LiveMathBench	100	35.55	5.00	31.93	0.00	26.88	0.00

Spurious Reward & Data Contamination

Table 2: Accuracy (Exact Match, EM) and ROUGE-L scores on several datasets (lower scores in gray) under different prompt prefix ratios in greedy decoding mode without applying chat template, namely ***Greedy (w/o Template)*** configuration.

Model	Dataset	Size	80%-Problem		60%-Problem		40%-Problem	
			RougeL	EM	RougeL	EM	RougeL	EM
Qwen2.5-Math-7B	MATH-500	500	81.25	**65.80**	78.06	**54.60**	69.01	**39.20**
	AMC	83	77.38	**55.42**	70.25	**42.17**	75.17	**36.14**
	AIME2024	30	74.04	**56.67**	55.31	20.00	57.72	16.67
	AIME2025	30	54.71	16.67	34.88	0.00	27.43	0.00
	MinervaMath	272	36.08	2.94	31.22	0.37	29.35	0.00
	LiveMathBench	100	42.76	5.00	32.78	0.00	29.97	0.00

RL Dataset	Dataset Size	MATH 500	AIME 2024	AMC 2023	Minerva Math	Olympiad-Bench	AIME 2025	Avg.
Qwen2.5-Math-7B [24] + GRPO								
NA	NA	$51.0_{+0.0}$	$12.1_{+0.0}$	$35.3_{+0.0}$	$11.0_{+0.0}$	$18.2_{+0.0}$	$6.7_{+0.0}$	$22.4_{+0.0}$
DSR-sub	1209	$78.6_{+27.6}$	$25.8_{+13.7}$	$62.5_{+27.2}$	$33.8_{+22.8}$	$41.6_{+23.4}$	$\mathbf{14.6}_{+7.9}$	$\mathbf{42.8}_{+20.4}$
$\{\pi_1\}$	1	$\mathbf{79.2}_{+28.2}$	$23.8_{+11.7}$	$60.3_{+25.0}$	$27.9_{+16.9}$	$39.1_{+20.9}$	$10.8_{+4.1}$	$40.2_{+17.8}$
$\{\pi_1, \pi_{13}\}$	2	$\mathbf{79.2}_{+28.2}$	$21.7_{+9.6}$	$58.8_{+23.5}$	$35.3_{+24.3}$	$40.9_{+22.7}$	$12.1_{+5.4}$	$41.3_{+18.9}$
$\{\pi_1, \pi_2, \pi_{13}, \pi_{1209}\}$	4	$78.6_{+27.6}$	$22.5_{+10.4}$	$61.9_{+26.6}$	$\mathbf{36.0}_{+25.0}$	$\mathbf{43.7}_{+25.5}$	$12.1_{+5.4}$	$42.5_{+20.1}$
Random	16	$76.0_{+25.0}$	$22.1_{+10.0}$	$\mathbf{63.1}_{+27.8}$	$31.6_{+20.6}$	$35.6_{+17.4}$	$12.9_{+6.2}$	$40.2_{+17.8}$
$\{\pi_1, \ldots, \pi_{16}\}$	16	$77.8_{+26.8}$	$\mathbf{30.4}_{+18.3}$	$62.2_{+26.9}$	$35.3_{+24.3}$	$39.9_{+21.7}$	$9.6_{+2.9}$	$42.5_{+20.1}$

Format reward baseline for Minerva and aime25: 24.3 & 6.7

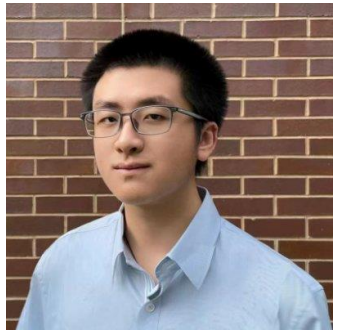
Spurious Reward & Data Contamination

- I think data contamination indeed happens, but it would not make 1-shot RLVR's conclusion fail.
- Mid-training or pretraining with open-source training data is really important for research.

Authors

Yiping Wang

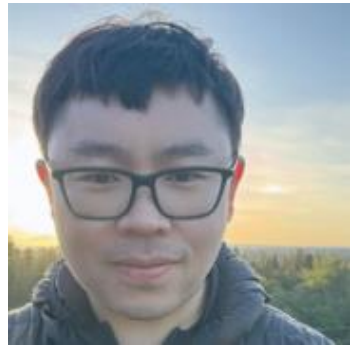
Qing Yang



Zhiyuan Zeng

Liliang Ren

Lucas Liu

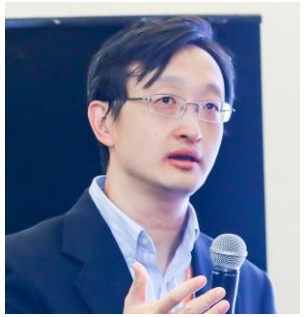


Baolin Peng

Hao Cheng

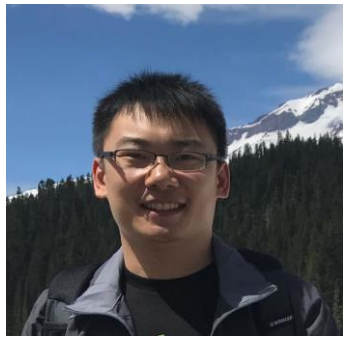
Xuehai He

Kuan Wang



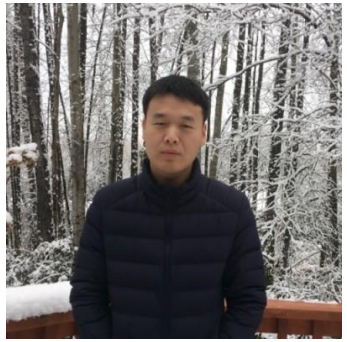
Jianfeng Gao

Weizhu Chen



Shuohang Wang

Simon S. Du



Yelong Shen

Thank You