Co-rewarding: Stable Self-supervised RL for Eliciting Reasoning in Large Language Models

Zizhuo Zhang*, Jianing Zhu*, Xinmu Ge*, Zihua Zhao*,
Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, Bo Han

TMLR Group, Hong Kong Baptist University
CMIC, Shanghai Jiao Tong University
Shanghai Innovation Institute
*Equal Contribution

Paper

Code

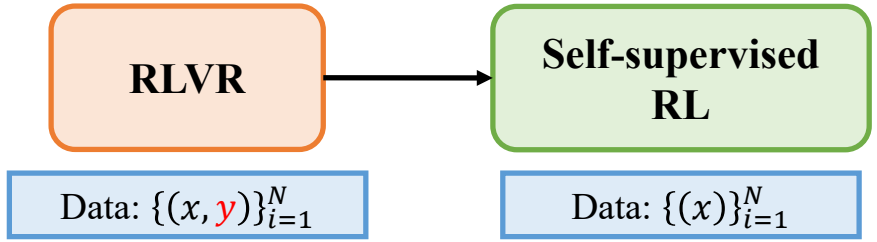
Model & Datasets

Zizhuo Zhang
cszzzhang@comp.hkbu.edu.hk

Main Contribution

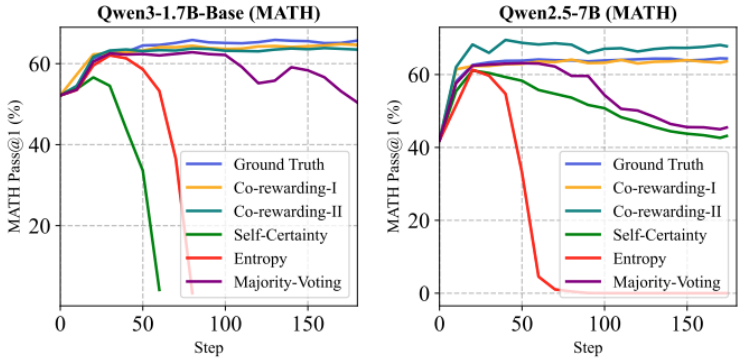
New Scenario: Label-free RL for LLM reasoning:

- We focus on **self-supervised RL** without requiring GT labels in LLM reasoning**,** compared to traditional RLVR with requiring GT labels as supervision.



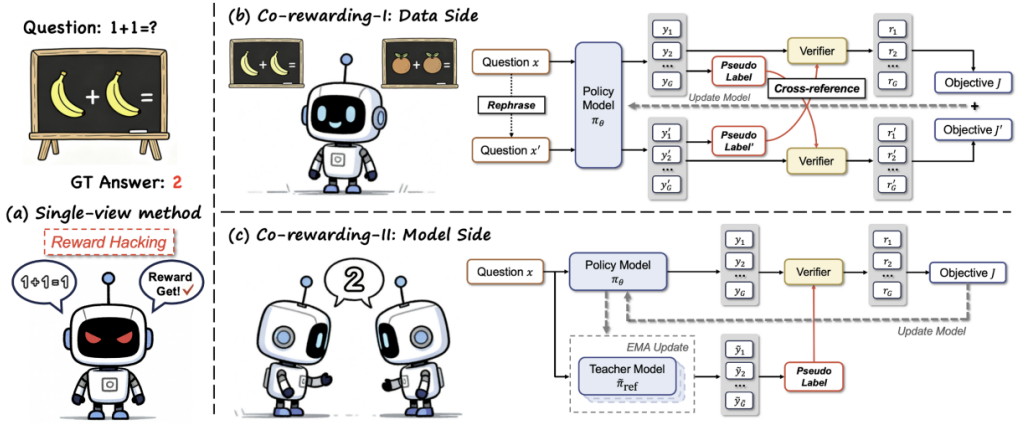
New Framework: Stable self-supervised RL framework:

- We propose **Co-rewarding**, a stable self-supervised RL framework that avoid training collapse issue that is frequently encountered in existing self-rewarding methods.



Multiple Instantiations of Co-rewarding framework:

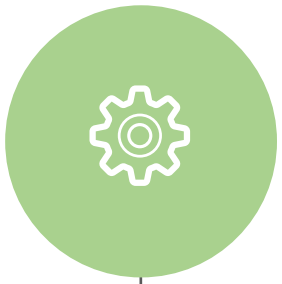
- We instantiate Co-rewarding in two ways from different perspectives
- **Co-rewarding-I:** a data-side instantiation that seeks reward signal from contrastive agreement across semantically analogous questions.
- **Co-rewarding-II:** a model-side instantiation that maintains a slowly-update teacher with pseudo labels to realize self-distillation.



Outline

Background
LLM Reasoning
RLVR

Related Work
Self-Certainty Maximization
Entropy Minimization
Majority-Voting



Motivation
Training Collapse
Reward Hacking

Methodology
Co-rewarding Framework
Co-rewarding-I: data side
Co-rewarding-II: model side

Experiments
Better Performance
Stable Training Process

Background: Large Language Model (LLM) Reasoning

LLM Reasoning aims to endow language models with the capacity to think, infer, and solve complex problems step-by-step in a manner resembling human reasoning.

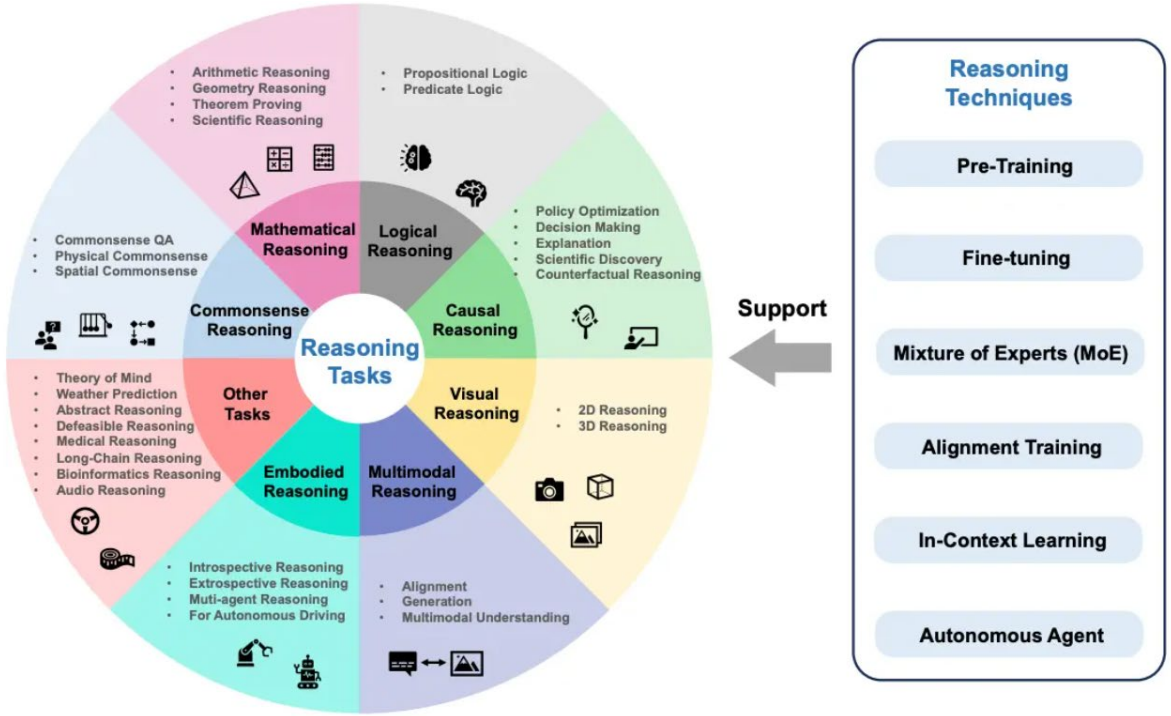


Figure refers to Sun et al.

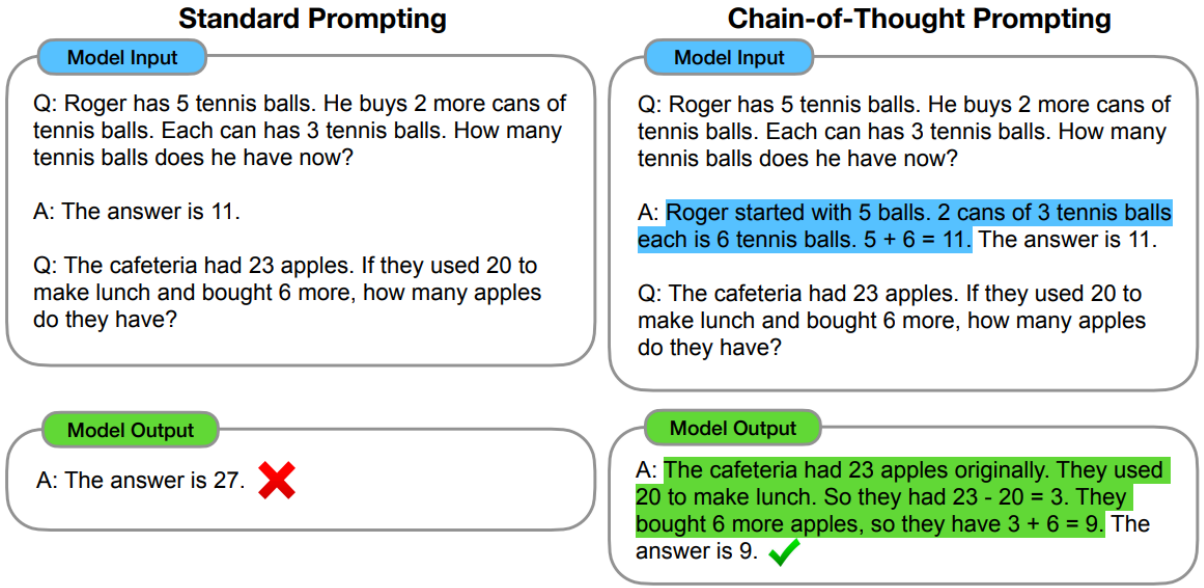


Figure refers to Wei et al.

Sun et al. Reasoning with Language Model Prompting: A Survey

Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Background: RLVR | GRPO

Reinforcement Learning with Verifiable Rewards (RLVR)

- RLVR is a paradigm that trains LLMs using rewards derived from objectively verifiable signals, such as correctness of math problems, execution of code.

Group Relative Policy Optimization (GRPO)

- GRPO is a variant RL algorithm of PPO, which foregoes the critic model, instead estimating the baseline from group scores.

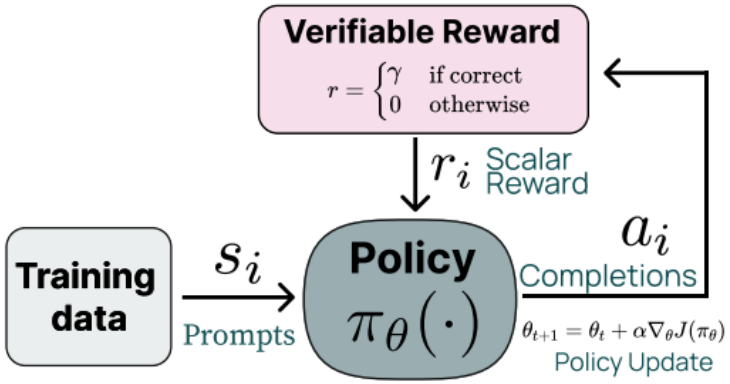


Figure refers to Tulu 3.

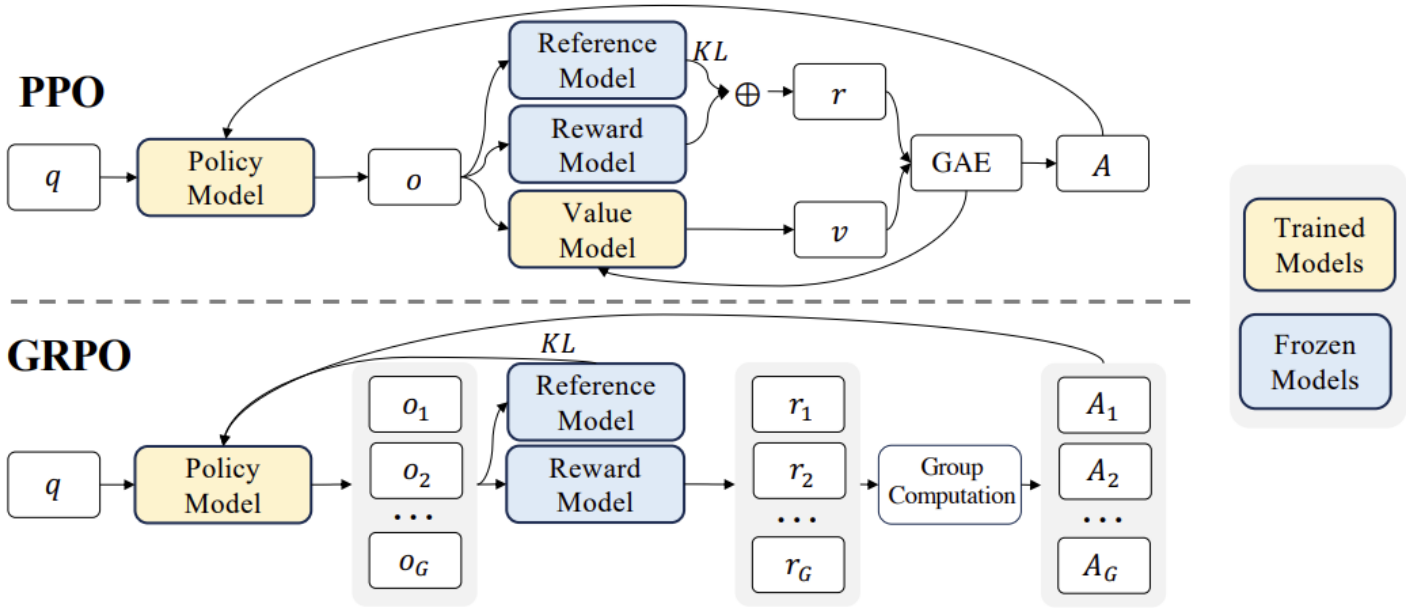


Figure refers to DeepSeekMath.

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models
Tulu 3: Pushing Frontiers in Open Language Model Post-Training
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

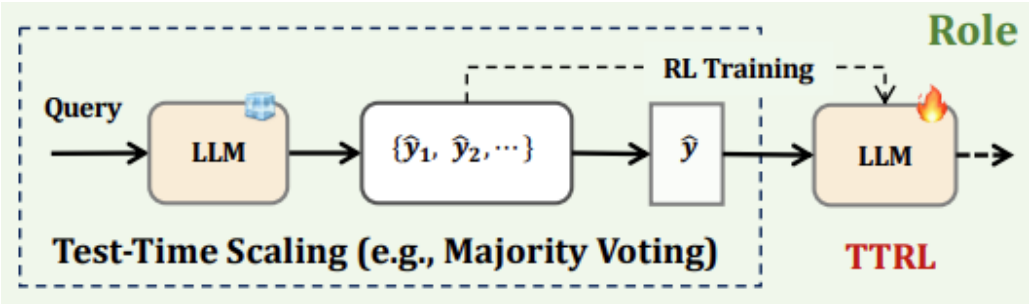
Related Work: Self-supervised RL Methods

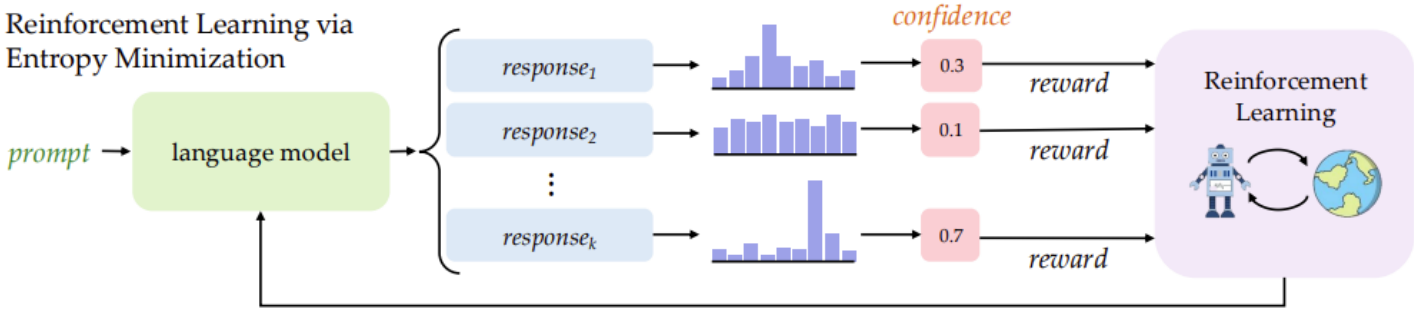
From RLVR to Self-supervised RL

- RLVR depends on **human-annotated labels (GT)** to verify the correctness of the answers of LLMs, which is a major bottleneck in the spirit of the scaling law.
- This motivates us to explore the **self-supervised RL** methods without reliance on GT labels.

Existing Self-supervised RL Methods:

- **Entropy-based:** Self-Certainty Maximization [1], Entropy Minimization [2,3], to enhance the confidence of the LLM.
- **Consensus-based:** Majority-Voting [4,5] to improve the consensus across multiple rollouts of the LLM for the same question.

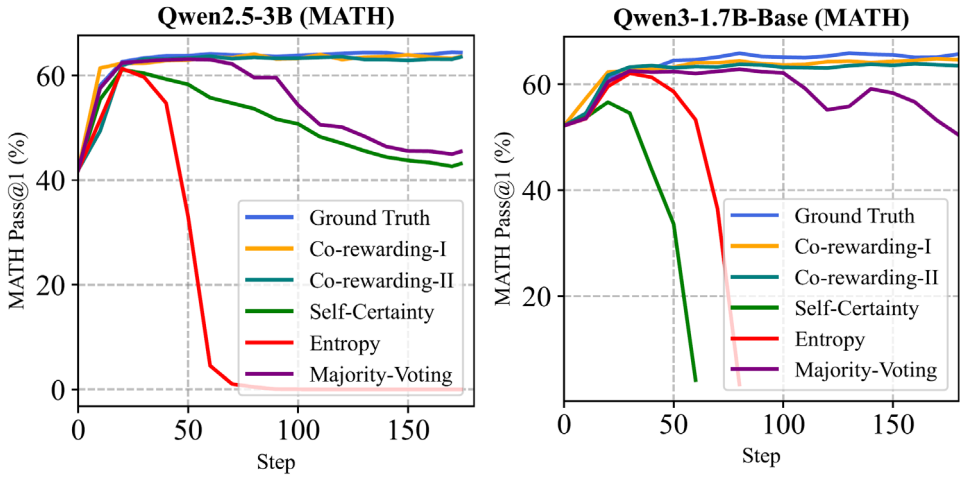


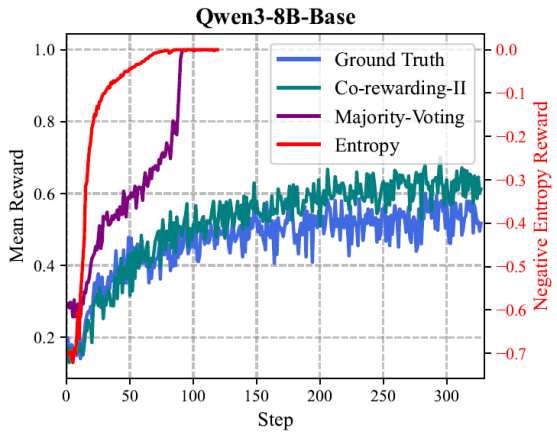


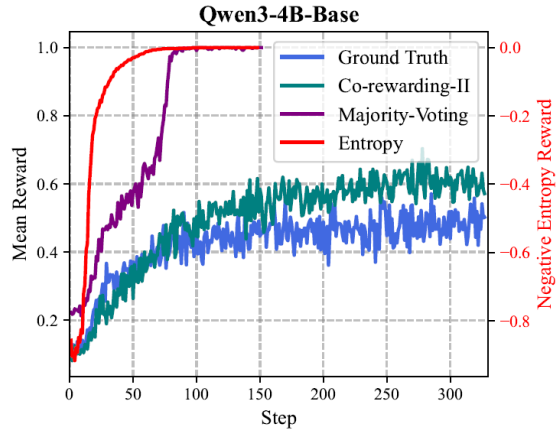
[1] Zhao et al. Learning to Reason without External Rewards
[2] Prabhudesai et al. Maximizing Confidence Alone Improves Reasoning
[3] Zhang et al. Right Question is Already Half the Answer: Fully Unsupervised LLM Reasoning Incentivization
[4] Zuo et al. TTRL: Test-Time Reinforcement Learning
[5] Shafayat et al. Can Large Reasoning Models Self-Train?

Motivation: Training Collapse | Reward Hacking

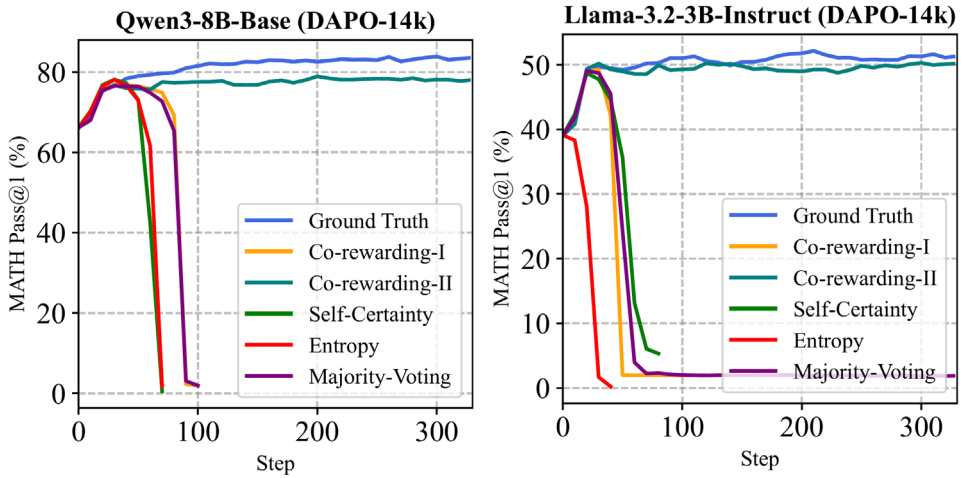
- Existing self-rewarding methods frequently encounter **training collapse** issue due to **reward hacking**.
- Existing single-view self-supervision signal easily forms the self-consistent illusion, yielding reward hacking.

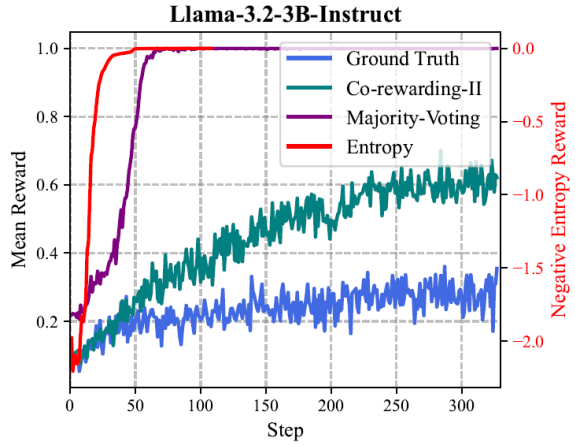


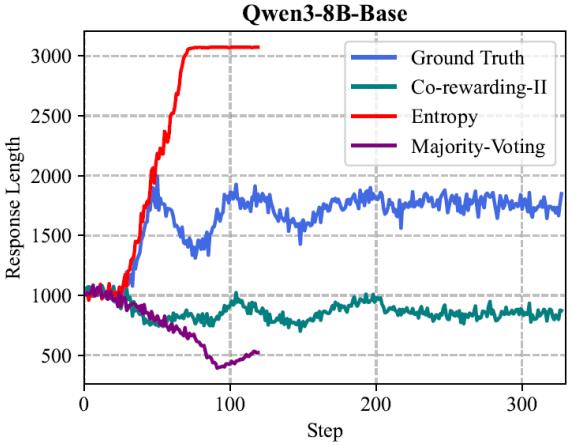




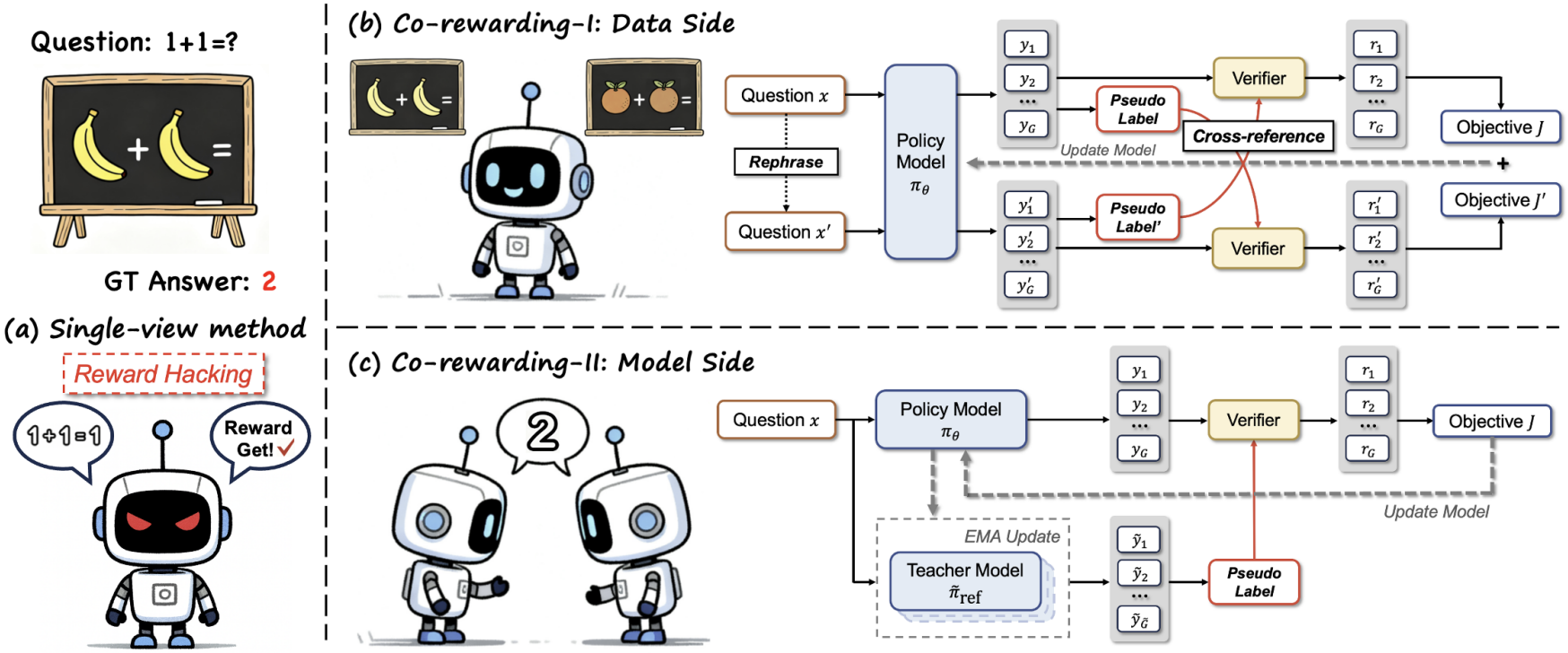
Training Collapse ← ***Result in*** → **Reward Hacking**





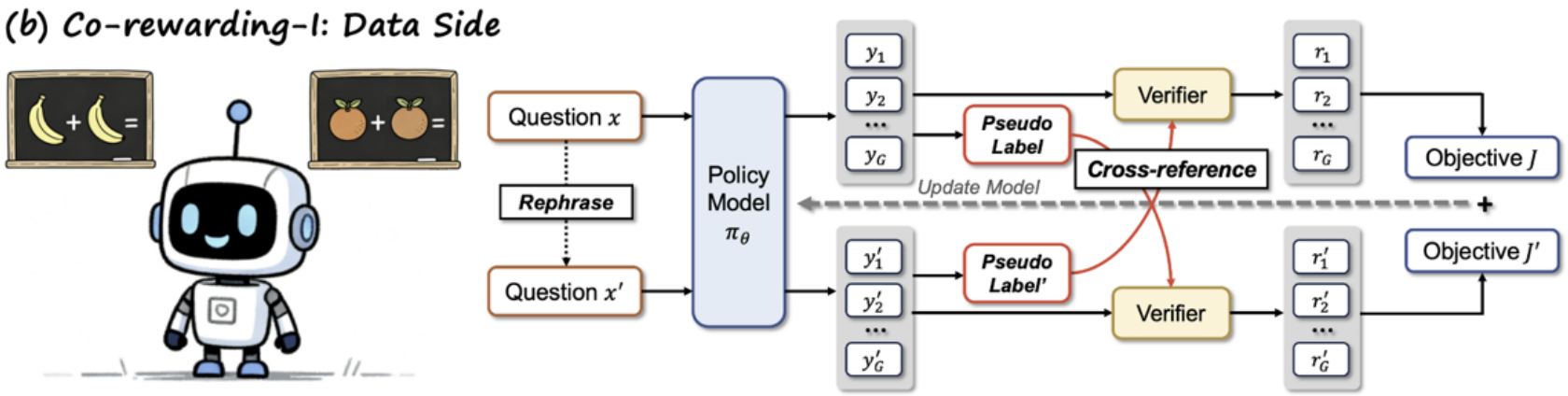


Methodology: *Co-rewarding Framework*



- ***Co-rewarding*** is a novel self-supervised RL framework that mitigates training collapse by seeking complementary supervision from another view.
- **Two types of instantiations:**
  - ***Co-rewarding-I:*** a data-side instantiation that derives reward signals from contrastive agreement across semantically analogous questions.
  - ***Co-rewarding-II:*** a model-side instantiation that maintains a slowly-updated teacher with pseudo labels to realize self-distillation to self-supervised quickly-learned student policy.

Co-rewarding-I: Data-side Instantiation



$$\mathcal{J}_{\text{Co-rewarding-I}}(\theta) = \underbrace{\mathbb{E}_{x\in\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\mathcal{R}_\theta(\hat{A})}_{\mathcal{J}_{\text{original}}(\theta)} + \underbrace{\mathbb{E}_{x'\in\mathcal{D}',\{y_i'\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x')}\mathcal{R}_\theta(\hat{A}')}_{\mathcal{J}_{\text{rephrased}}(\theta)}, \tag{6}$$

$$\hat{A}_i = \frac{r(y_{\text{v}}', y_i) - \text{mean}(\{r(y_{\text{v}}', y_i)\}_{i=1}^{G})}{\text{std}(\{r(y_{\text{v}}', y_i)\}_{i=1}^{G})}, \; \hat{A}_i' = \frac{r(y_{\text{v}}, y_i') - \text{mean}(\{r(y_{\text{v}}, y_i')\}_{i=1}^{G})}{\text{std}(\{r(y_{\text{v}}, y_i')\}_{i=1}^{G})}. \tag{7}$$

$$y_{\text{v}} \leftarrow \arg\max_{y*} \sum_{i=1}^{G} 1[\text{ans}(y_i) = \text{ans}(y*)], \quad y_{\text{v}}' \leftarrow \arg\max_{y*} \sum_{i=1}^{G} 1[\text{ans}(y_i') = \text{ans}(y*)]. \tag{8}$$

Algorithm 1 *Co-rewarding-I*

Input: policy model π_θ, learning rate η, training dataset $\mathcal{D}$, rephrased training dataset $\mathcal{D}'$, total iterations K.
Output: trained policy model π_θ.
for all iteration $k = 1, \ldots, K$ **do**
  Sample mini-batch inputs $\mathcal{B} \subseteq \mathcal{D}$ and $\mathcal{B}' \subseteq \mathcal{D}'$.
  for all input question $x \in \mathcal{B}$ and $x' \in \mathcal{B}'$ **do**
    Sample rollouts $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$.
    Sample rollouts $\{y_i'\}_{i=1}^{G'} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x')$.
    Obtain pseudo labels by Eq. (8).
    Estimate relative advantages by Eq. (7).
    Compute the objective by Eq. (6).
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{J}_{\text{Co-rewarding-I}}(\theta)$.
  end for
end for

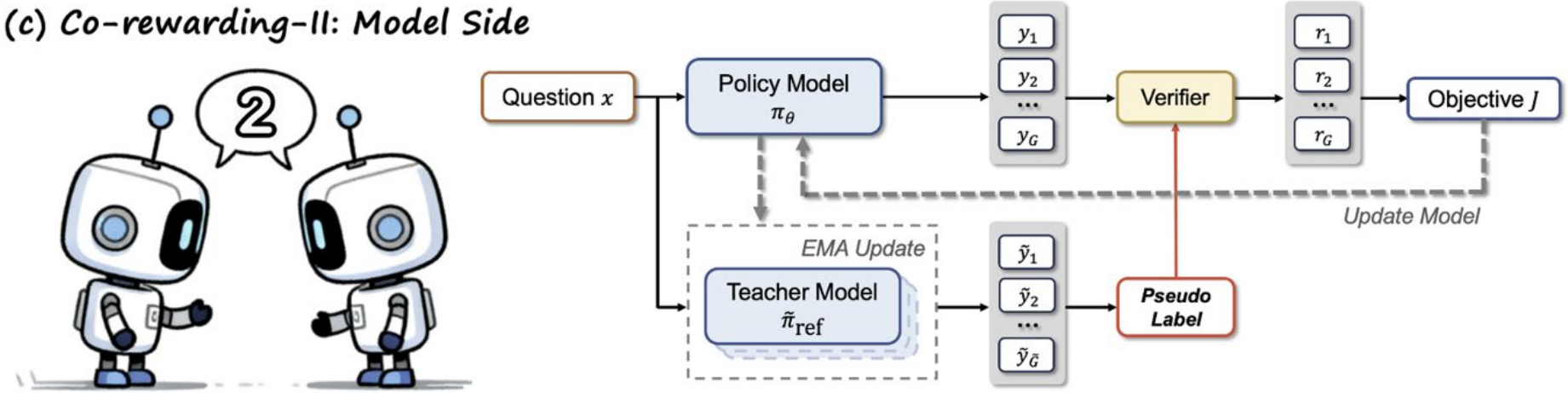
- Questions that **share the same mathematical essence but differ in surface form** (e.g., via paraphrasing, or reformatting) should elicit the comparably valid and similar reasoning results.
- ***Co-rewarding-I* defines contrastive agreement** as a principle that aligns model reasoning outputs, treating consistent inter-view agreement as a signal for valid inference.

Co-rewarding-I: Data-side Instantiation

Original Question	**Rephrased Question**
Sam is hired for a 20-day period. On days that he works, he earns $60. For each day that he does not work, $30 is subtracted from his earnings. At the end of the 20-day period, he received $660. How many days did he not work?	A contractor agrees to a job lasting 20 days. For every day the job is completed on time, the contractor earns $60. However, for each day the work is delayed, a fine of $30 is applied. After the 20-day period, the contractor's total earnings are $660. How many days was the job delayed?
Karen drove continuously from 9:40 a.m. until 1:20 p.m. of the same day and covered a distance of 165 miles. What was her average speed in miles per hour?	A traveler set off at 9:40 a.m. and reached their destination at 1:20 p.m. the same day after traveling a total of 165 miles. What was their average speed during the trip in miles per hour?
Solve for x: $\frac{1}{2} + \frac{1}{x} = \frac{5}{6}$.	A tank is partially filled by two different pipes. One pipe fills half the tank in an hour, and together with another pipe, they fill five-sixths of the tank in the same time. If the second pipe alone fills $\frac{1}{x}$ of the tank in an hour, find the value of x.

- Rephrasing does not alter the mathematical essence of the questions.
- We assume that both original question and rephrased question should have similar reasoning results.
- **We use pseudo-labels generated from each to supervise the other.**

Co-rewarding-II: Model-side Instantiation



$$\mathcal{J}^{(k)}_{\text{Co-rewarding-II}}(\theta) = \mathbb{E}_{x\in\mathcal{D},\underbrace{\{y_i\}_{i=1}^{G}\sim\pi^{(k)}_{\theta_{\text{old}}}(\cdot|x)}_{\text{policy student rollout}},\underbrace{\{\tilde{y}_j\}_{j=1}^{\tilde{G}}\sim\tilde{\pi}^{(k)}_{\text{ref}}(\cdot|x)}_{\text{reference teacher rollout}}} \mathcal{R}_\theta(\hat{A}^{(k)}), \tag{9}$$

$$\hat{A}^{(k)}_i = \frac{r(\tilde{y}^{(k)}_{\text{v}}, y_i) - \text{mean}(\{r(\tilde{y}^{(k)}_{\text{v}}, y_i)\}_{i=1}^{G})}{\text{std}(\{r(\tilde{y}^{(k)}_{\text{v}}, y_i)\}_{i=1}^{G})}, \quad \tilde{y}^{(k)}_{\text{v}} = \arg\max_{y^*} \sum_{j=1}^{\tilde{G}} \mathbf{1}[\text{ans}(\tilde{y}_j) = \text{ans}(y^*)], \tag{10}$$

$$\tilde{\pi}^{(k)}_{\text{ref}} \leftarrow \alpha^{(k)} \cdot \tilde{\pi}^{(k-1)}_{\text{ref}} + (1-\alpha^{(k)}) \cdot \pi^{(k)}_{\theta_{\text{old}}}, \quad \alpha^{(k)} = 1 - \frac{(\alpha_{\text{end}} - \alpha_{\text{start}})}{2}\left(1 + \cos\left(\frac{\pi k}{K}\right)\right) \tag{11}$$

Algorithm 2 *Co-rewarding-II*

Input: policy model π_θ, learning rate η, training dataset $\mathcal{D}$, total iterations K.
Output: trained policy model π_θ.
for iteration $k = 1, \ldots, K$ **do**
  Sample mini-batch $\mathcal{B} \subseteq \mathcal{D}$.
  for all $x \in \mathcal{B}$ **do**
    Sample rollouts $\{y_i\}_{i=1}^{G} \sim \pi^{(k)}_{\theta_{\text{old}}}(\cdot \mid x)$.
    Update the reference teacher by Eq. (11).
    Sample rollouts $\{\tilde{y}_j\}_{j=1}^{\tilde{G}} \sim \tilde{\pi}^{(k)}_{\text{ref}}(\cdot \mid x)$.
    Obtain pseudo label from $\{\tilde{y}_j\}_{j=1}^{\tilde{G}}$ by Eq. (10).
    Estimate the relative advantage by Eq. (10).
    Compute the objective by Eq. (9).
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{J}^{(k)}_{\text{Co-rewarding-II}}(\theta)$.
  end for
end for

- ***Co-rewarding-II*** sources pseudo-labels from a teacher reference, which disentangle the self-supervision reward from the online policy.
- **The teacher is dynamically updated as an exponential moving average (EMA)** of the student policy to ensure pseudo-label quality improving as the policy improves.

Experiment Setup

- **Multiple LLM Families:**
  - **Qwen2.5 Series:** Qwen2.5-3B, Qwen2.5-7B
  - **Qwen3 Series:** Qwen3-1.7B-Base, Qwen3-4B-Base, Qwen3-8B-Base
  - **Llama Series:** Llama-3.2-3B-Instruct

- **Multiple Training Sets:**
  - **MATH:** 7,500 training samples, DigitalLearningGmbH/MATH-lighteval
  - **DAPO-14k:** 14.1k training samples, open-r1/DAPO-Math-17k-Processed
  - **OpenRS:** 7,000 training samples, knoveleng/open-rs

- **Baselines:**
  - **Self-Certainty [1]:** Self-Certainty maximization
  - **Entropy [2]:** Entropy minimization
  - **Majority-Voting [3]:** Pursue answer-level consensus via majority-voting from multiple rollouts

- **Evaluation Benchmark:**
  - **Mathematical Reasoning:** MATH500, GSM8K, AMC
  - **Code Generation:** LiveCodeBench, CRUX
  - **General Domain:** MMLU-Pro, IFEval

[1] Zhao et al. Learning to Reason without External Rewards
[2] Prabhudesai et al. Maximizing Confidence Alone Improves Reasoning
[3] Shafayat et al. Can Large Reasoning Models Self-Train?

Overall Performance

MATH Training Set

Training Set: MATH	Mathematics				Code		Instruction	Multi-Task
Methods	**MATH500**	**GSM8K**	**AMC**	**AIME24**	**LiveCode**	**CRUX**	**IFEval**	**MMLU-Pro**
Qwen3-8B-Base								
Before RL	72.4	27.82	20.93	3.75	23.41	54.75	50.89	52.92
- GT-Reward (Shao et al., 2024)	82.6	87.26	54.22	17.15	30.52	63.25	52.78	57.11
- Self-Certainty (Zhao et al., 2025b)	80.2	80.74	50.75	15.73	27.20	64.38	50.98	54.17
- Entropy (Prabhudesai et al., 2025)	80.2	87.19	49.54	15.63	29.38	62.00	51.81	54.86
- Majority-Voting (Shafayat et al., 2025)	79.8	89.76	49.09	15.83	30.52	63.38	51.80	56.93
- Co-rewarding-I (Ours)	81.2	93.70	51.20	15.10	30.81	66.00	55.79	59.95
- Co-rewarding-II (Ours)	80.8	92.42	53.46	14.48	30.23	62.83	60.70	57.50
- Co-rewarding-III (Ours)	81.4	90.98	54.07	13.33	30.71	63.75	53.69	59.10
Qwen3-4B-Base								
Before RL	71.2	26.15	21.08	4.58	11.00	38.88	46.43	47.23
- GT-Reward (Shao et al., 2024)	78.6	89.76	51.20	15.00	26.07	55.38	47.80	53.96
- Self-Certainty (Zhao et al., 2025b)	71.6	71.79	38.86	11.67	22.37	57.00	48.15	48.93
- Entropy (Prabhudesai et al., 2025)	77.0	88.10	47.44	10.94	25.59	52.88	50.44	49.90
- Majority-Voting (Shafayat et al., 2025)	77.4	90.07	45.33	10.10	26.54	57.50	48.78	54.35
- Co-rewarding-I (Ours)	78.8	91.28	46.08	13.85	26.64	56.50	50.35	53.26
- Co-rewarding-II (Ours)	78.0	88.86	45.93	12.17	26.25	55.00	51.30	53.88
- Co-rewarding-III (Ours)	78.6	90.75	48.80	12.71	26.16	56.00	49.23	53.08
Llama-3.2-3B-Instruct								
Before RL	39.2	65.73	10.54	3.75	9.86	25.37	57.32	31.14
- GT-Reward (Shao et al., 2024)	47.0	77.94	22.14	11.67	9.57	31.87	47.51	34.32
- Self-Certainty (Zhao et al., 2025b)	43.4	74.91	18.83	6.88	9.95	25.87	54.88	33.34
- Entropy (Prabhudesai et al., 2025)	43.4	66.19	20.18	6.56	11.66	24.62	54.70	33.52
- Majority-Voting (Shafayat et al., 2025)	46.8	78.77	20.48	9.27	11.00	31.25	47.96	33.18
- Co-rewarding-I (Ours)	50.2	79.45	23.80	10.00	11.28	29.88	48.89	33.77
- Co-rewarding-II (Ours)	49.8	79.30	22.59	10.73	10.80	30.63	49.90	33.61
- Co-rewarding-III (Ours)	51.6	79.91	25.45	10.42	10.43	32.50	46.37	34.50

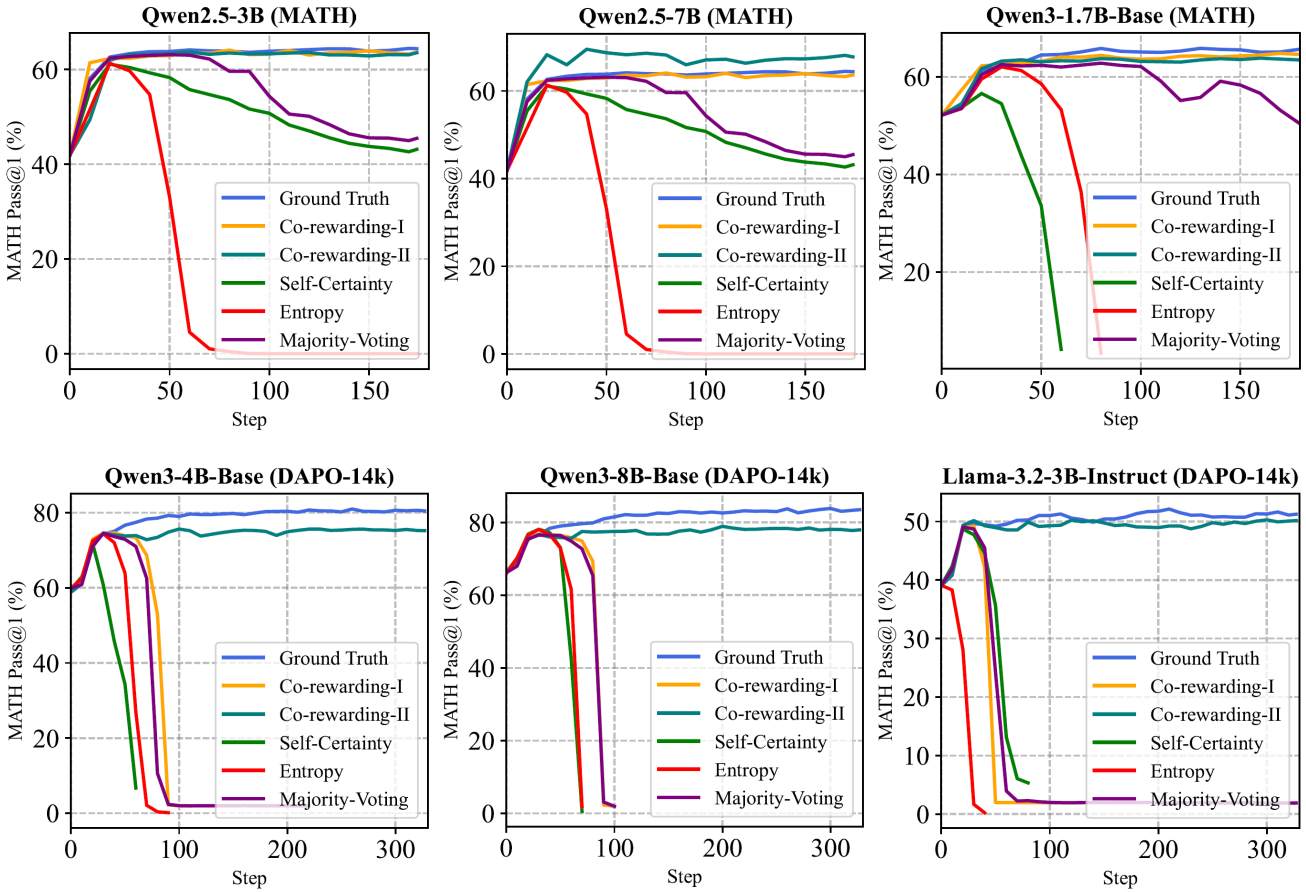
DAPO-14k Training Set

Training Set: DAPO-14k	Mathematics				Code		Instruction	Multi-Task
Methods	**MATH500**	**GSM8K**	**AMC**	**AIME24**	**LiveCode**	**CRUX**	**IFEval**	**MMLU-Pro**
Qwen3-8B-Base								
Before RL	72.4	27.82	20.93	3.75	23.41	54.75	50.89	52.92
- GT-Reward (Shao et al., 2024)	86.6	87.19	61.75	24.58	30.52	63.75	53.11	60.27
- Self-Certainty (Zhao et al., 2025b)	82.0	77.63	49.85	15.00	27.77	60.75	50.58	54.24
- Entropy (Prabhudesai et al., 2025)	79.4	80.82	45.48	15.00	30.14	62.00	51.56	54.57
- Majority-Voting (Shafayat et al., 2025)	78.6	91.66	50.00	11.25	30.33	61.62	51.54	55.65
- Co-rewarding-I (Ours)	78.4	88.02	51.20	11.88	29.38	62.50	50.17	55.39
- Co-rewarding-II (Ours)	80.6	94.01	54.37	16.35	31.66	67.12	53.31	59.83
- Co-rewarding-III (Ours)	81.6	92.27	53.77	17.71	32.70	66.75	55.85	60.02
Qwen3-4B-Base								
Before RL	71.2	26.15	21.08	4.58	11.00	38.88	46.43	47.23
- GT-Reward (Shao et al., 2024)	83.6	85.14	52.86	20.63	18.58	56.88	47.70	55.35
- Self-Certainty (Zhao et al., 2025b)	68.4	44.81	35.39	8.85	25.88	50.12	45.58	48.84
- Entropy (Prabhudesai et al., 2025)	76.6	82.79	43.37	12.81	26.35	50.75	48.20	50.22
- Majority-Voting (Shafayat et al., 2025)	73.4	64.06	40.81	9.17	26.16	53.00	48.91	51.06
- Co-rewarding-I (Ours)	73.8	75.89	43.83	10.63	26.25	50.12	46.84	51.51
- Co-rewarding-II (Ours)	77.8	91.89	48.49	14.27	26.64	54.87	48.90	52.83
- Co-rewarding-III (Ours)	79.2	90.45	48.95	15.10	27.58	54.87	50.30	54.79
Llama-3.2-3B-Instruct								
Before RL	39.2	65.73	10.54	3.75	9.86	25.37	57.32	31.14
- GT-Reward (Shao et al., 2024)	49.4	78.17	25.90	9.17	10.33	31.37	53.10	33.83
- Self-Certainty (Zhao et al., 2025b)	42.4	74.71	17.32	4.79	11.18	28.38	54.50	33.51
- Entropy (Prabhudesai et al., 2025)	44.0	65.85	17.32	6.56	9.95	25.00	55.78	31.95
- Majority-Voting (Shafayat et al., 2025)	42.8	70.96	17.62	8.74	10.14	29.50	54.07	32.95
- Co-rewarding-I (Ours)	46.0	70.58	20.93	7.08	9.57	27.25	53.04	32.61
- Co-rewarding-II (Ours)	49.8	78.62	19.73	8.02	10.43	32.25	51.92	34.46
- Co-rewarding-III (Ours)	48.6	76.95	21.84	8.13	9.86	30.50	49.92	34.01

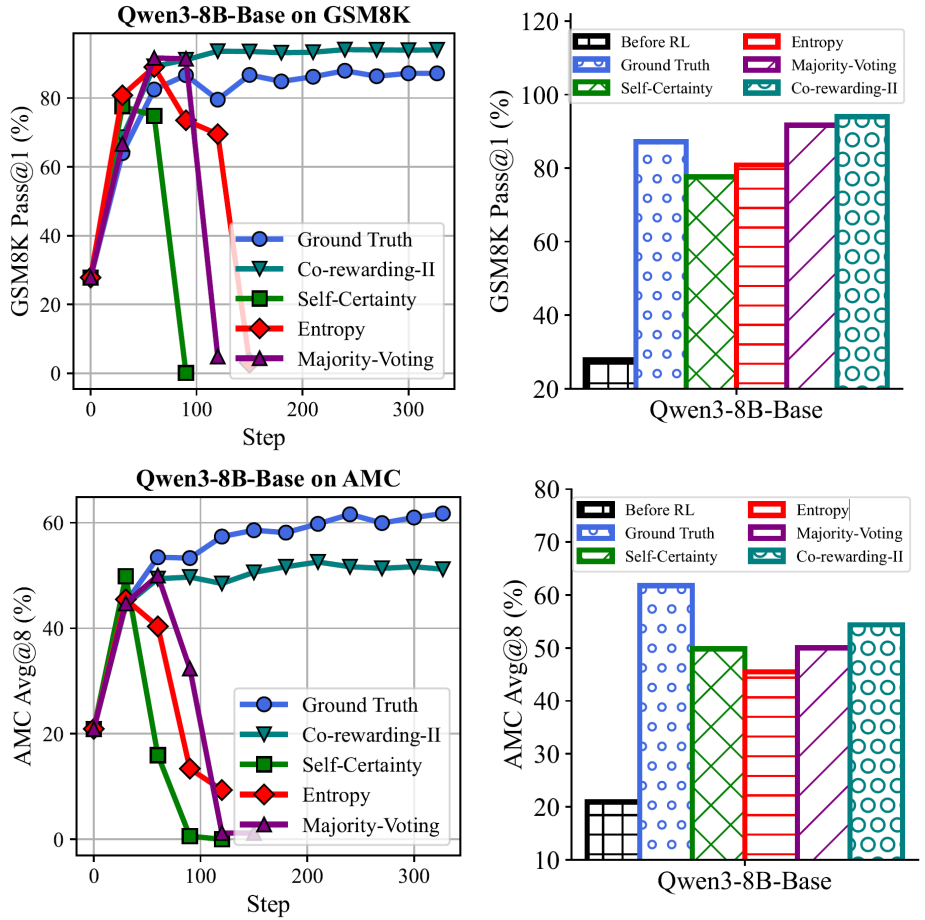
- **Superior Performance of Co-rewarding over self-rewarding baselines.** Co-rewarding-I achieves an average relative performance gain of **+3.46%** over the best baselines across three mathematical benchmarks and models in Table 1, while Co-rewarding-II achieves a larger average relative gain of **+7.29%** in Table 2.
- **Surpassing GT-Reward in certain benchmarks.** Co-rewarding-II achieves a remarkably high Pass@1 of **94.01%** with Qwen3-8B-Base on GSM8K.
- **Code generalization with preserved general performance.**

Stability to Mitigate Training Collapse

Validation Performance Curve



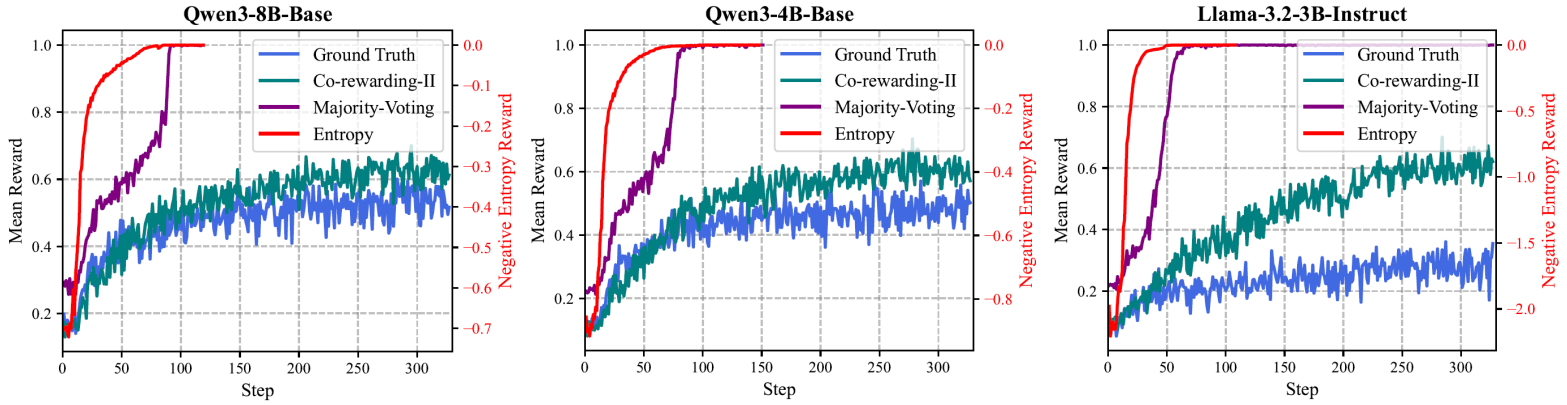
Performance on GSM8K and AMC



Observation: Co-rewarding alleviates collapse and provides stable self-supervised RL.

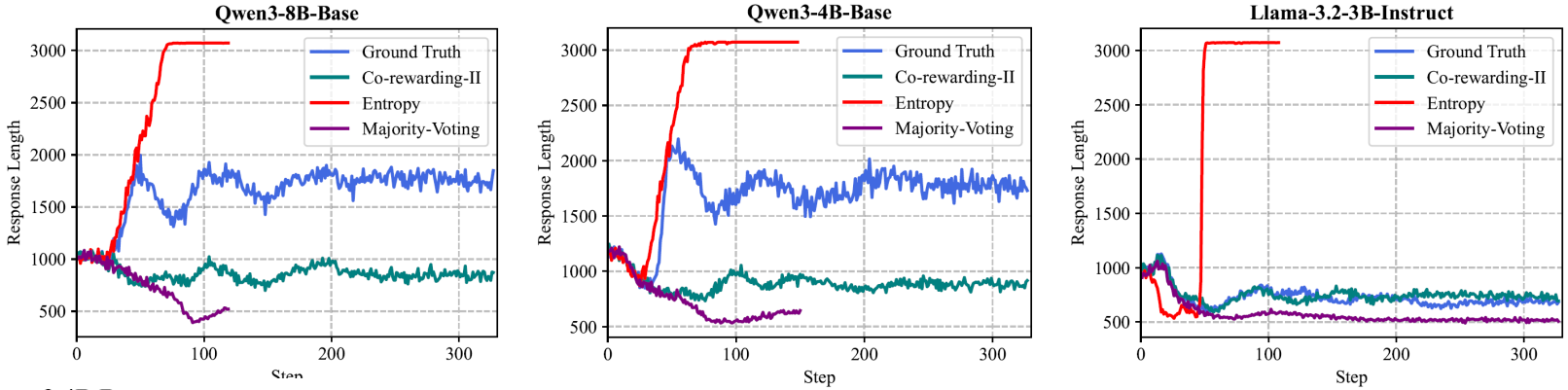
Observation: Importance of stability for further performance gain.

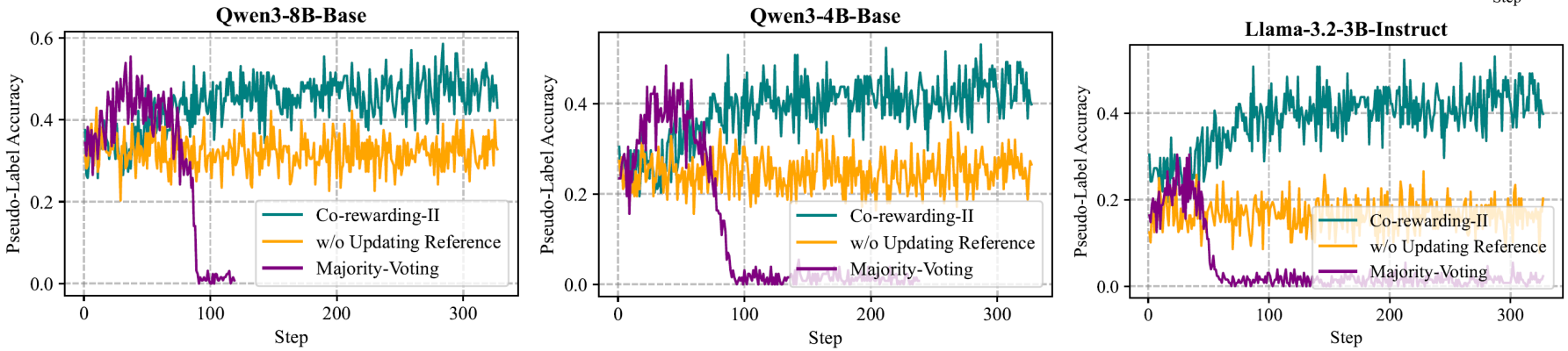
Reward | Response Length | Pseudo Label Accuracy



Observation: Co-rewarding avoids reward hacking and provides stable self-supervised RL process.

Observation: Co-rewarding attempts to balance exploration-exploitation.





Observation: EMA is essential in Co-rewarding-II for improving pseudo-label quality.

Ablation Study

Training Set	Methods	MATH500	GSM8K	AMC	AIME24	LiveCode	CRUX	IFEval	MMLU-Pro
					Qwen3-8B-Base				
MATH	Co-rewarding-I	**81.2**	**93.70**	**51.20**	15.10	30.81	**66.00**	**55.79**	**59.95**
	- Majority-Voting w/ Union	80.2	93.48	49.70	15.63	**31.94**	64.88	54.25	59.51
	- Majority-Voting w/ Original	79.8	89.76	49.09	**15.83**	30.52	63.38	51.80	56.93
	- Majority-Voting w/ Rephrased	79.2	91.51	50.75	14.17	31.66	60.38	52.24	57.26
	Co-rewarding-II	**80.8**	**92.42**	**53.46**	**14.48**	30.23	**62.83**	**60.70**	**57.50**
	- w/o Updating Reference	79.2	89.46	51.51	13.96	**30.62**	61.75	56.93	51.85
					Llama-3.2-3B-Instruct				
	Co-rewarding-I	**50.2**	79.45	**23.80**	**10.00**	**11.28**	29.88	**48.89**	**33.77**
	- Majority-Voting w/ Union	48.0	**80.52**	21.84	9.69	10.14	30.00	43.35	34.05
	- Majority-Voting w/ Original	46.8	78.77	20.48	9.27	11.00	**31.25**	47.96	33.18
	- Majority-Voting w/ Rephrased	44.0	78.85	21.23	8.85	10.04	17.25	47.84	33.72
	Co-rewarding-II	**49.8**	**79.30**	**22.59**	**10.73**	**10.80**	30.63	**49.90**	**33.61**
	- w/o Updating Reference	47.0	78.92	22.29	9.06	5.50	**31.25**	47.88	33.32
					Qwen3-8B-Base				
DAPO-14k	Co-rewarding-II	**80.6**	**94.01**	**54.37**	**16.35**	**31.66**	**67.12**	**53.31**	**59.83**
	- w/o Updating Reference	78.0	88.40	51.66	15.94	30.62	63.75	52.48	58.01
					Llama-3.2-3B-Instruct				
	Co-rewarding-II	**49.8**	**78.62**	**19.73**	**8.02**	**10.43**	**32.25**	**51.92**	**34.46**
	- w/o Updating Reference	45.0	76.72	17.92	**8.02**	10.05	30.63	51.33	33.94

Observation: Each part contributes to Co-rewarding.

- For Co-rewarding-I, the quality of original and rephrased data is similar, while the cross supervision across views provides the crucial gains.
- For Co-rewarding-II, removing the EMA update of the reference teacher model causes clear degradation, highlighting the necessity of teacher updates for improving pseudo-label quality.

Case Study

Question: $\frac{m}{n}$ **is the Irreducible fraction value of** $\left[3+\frac{1}{3+\frac{1}{3+\frac{1}{3}}}\right]$**, what is the value of** $m+n$**?** **Correct Answer: 142**

To find the value of …, we need simplify the expression step by step.
First, let's simplify the innermost fraction:

$$3+\frac{1}{3}=3+\frac{1}{3}=\frac{9}{3}+\frac{1}{3}=\frac{10}{3}$$

Now, substitute this back into the expression:

……

Next, simplify the fraction: ……
Now, substitute this back into the expression:

……

To add these, we need a common denominator:

……

So, the irreducible fraction value of the expression is $\frac{109}{3}$. Here, $m = 109$ and $n = 33$.
Therefore, the value of $m + n$ is:

$$m + n = 109 + 33 = 142$$

Thus, the value of $m + n$ is 0.

Majority-Voting ✗

To find the value of $\frac{m}{n}$ ……
Let's simplify the expression step by step.
1. Simplify the innermost fraction:

$$3+\frac{1}{3}$$

……

2. Simplify the next fraction:

$$3+\frac{1}{3+\frac{1}{3}}=3+\frac{1}{\frac{10}{3}}$$

……

5. Simplify the next fraction:

$$3+\cfrac{1}{3+\cfrac{1}{3+\cfrac{1}{3+\cfrac{1}{3+\frac{1}{3}}}}}$$

Entropy *(repeat until max response)* ✗

To solve the problem, we need to simplify the given expression step by step and then find the irreducible fraction form.
First, we simplify the innermost fraction:

……

Now substitute this back into the expression:

……

Next, simplify the fraction inside the denominator:

……

Therefore, m = 109 and n = 33, and the value of (m + n) is: 109 + 33 = 142.
So, the final answer is: **142.**

Co-rewarding-I ✓

To solve the given expression ……, we need to simplify it step by step.

First, simplify the innermost fraction: ……

Next, substitute this back into the next fraction: …..

Now, simplify the expression $3+\frac{10}{33}$: ……

The fraction $\frac{109}{33}$ is already in its simplest form because ……
Therefore, (m + n = 109 + 33 = 142).

The value of (m + n) is **142.**

Co-rewarding-II ✓

- Majority-Voting exhibits reward hacking by boxing an incorrect answer "0" to pursue consensus.
- Entropy produces repetitive outputs as its decoding probability distribution collapses onto a narrow set of tokens during entropy minimization.
- Co-rewarding generates coherent step-by-step reasoning and correctly boxes the final answer.

Take-home Message

- We explore the **self-supervised RL paradigm, which eliminates the need for ground-truth (GT) labels** required by RLVR.
- We investigate how existing self-rewarding methods **suffer from training collapse as a consequence of reward hacking**.
- We propose **Co-rewarding, a novel self-supervised RL framework** that is initiated by the data and model sides to construct self-generate rewards to promote stably reasoning elicitation .
- We empirically demonstrate the effectiveness of Co-rewarding to achieve superior and stable reasoning performance on LLMs.

Thank you!

If you have any question, feel free to contact me.

Zizhuo Zhang
cszzzhang@comp.hkbu.edu.hk