

CO-REWARD: SELF-SUPERVISED REINFORCEMENT LEARNING FOR LARGE LANGUAGE MODEL REASONING VIA CONTRASTIVE AGREEMENT

Zizhuo Zhang^{1*} Jianing Zhu^{1*} Xinmu Ge^{2,3*} Zihua Zhao^{2*} Zhanke Zhou¹
 Xuan Li¹ Xiao Feng¹ Jiangchao Yao^{2†} Bo Han^{1†}

¹TMLR Group, Hong Kong Baptist University

²CMIC, Shanghai Jiao Tong University ³Shanghai Innovation Institute

ABSTRACT

Although reinforcement learning with verifiable rewards (RLVR) shows promise in improving the reasoning ability of large language models (LLMs), the scaling up dilemma remains due to the reliance on human annotated labels especially for complex tasks. Recent alternatives that explore various self-reward signals exhibit the eliciting potential of LLM reasoning, but suffer from the non-negligible collapse issue. Inspired by the success of self-supervised learning, we propose *Co-Reward*, a novel RL framework that leverages contrastive agreement across semantically analogical questions as a reward basis. Specifically, we construct a similar question for each training sample (without labels) and synthesize their individual surrogate labels through a simple rollout voting, and then the reward is constructed by cross-referring the labels of each question pair to enforce the internal reasoning consistency across analogical inputs. Intuitively, such a self-supervised reward-shaping mechanism increases the difficulty of learning collapse into a trivial solution, and promotes stable reasoning elicitation and improvement through expanding the input sample variants. Empirically, Co-Reward achieves superior performance compared to other self-reward baselines on multiple reasoning benchmarks and LLM series, and reaches or even surpasses ground-truth (GT) labeled reward, with improvements of up to +6.8% on MATH500 over GT reward on Llama-3.2-3B-Instruct. Our code is publicly available at <https://github.com/tmlr-group/Co-Reward>.

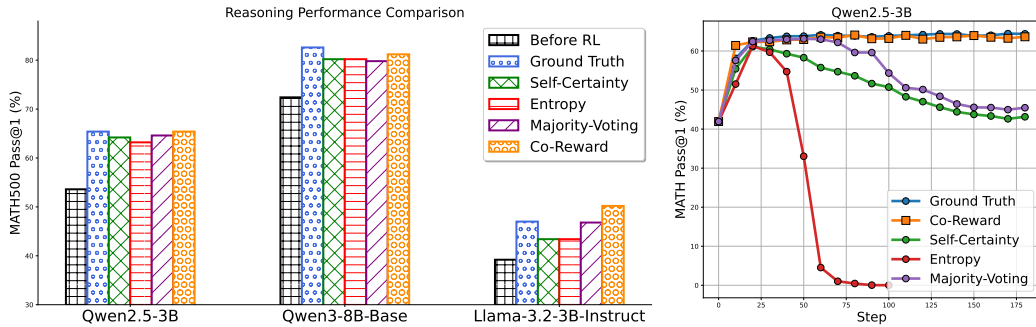


Figure 1: Performance overview. Left: the superiority of Co-Reward demonstrated by reasoning performance on MATH500 (Lightman et al., 2024) using different base models; Right: model collapse illustration with the significant drop of validation curves on baselines, also without ground truth.

1 INTRODUCTION

Large language models (LLMs) (Achiam et al., 2023; Dubey et al., 2024; Qwen et al., 2025) have demonstrated remarkable general-purpose capabilities in a wide range of linguistic tasks (Hendrycks

*Equal Contribution. † Correspondence to Bo Han (bhanml@comp.hkbu.edu.hk) and Jiangchao Yao (Sunarker@sjtu.edu.cn).

et al.). To enhance reasoning ability in specific domains, reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; Yu et al., 2025) is developed as an essential paradigm for post-training with externally verifiable signals, such as program execution results (Luo et al., 2025) or mathematical equivalence (Shao et al., 2024). Although performance improvement is impressive, the reliance on high-quality ground-truth supervision is considered a major obstacle (Ouyang et al., 2022; Bai et al., 2022) to scale up, which motivates the emerging exploration of other promising paradigms such as self-evolving (Wang et al., 2022) to refine reasoning behavior using only unlabeled data.

One prominent direction in this space involves leveraging self-reward signals (e.g., entropy information (Prabhudesai et al., 2025) and certainty (Zhao et al., 2025b)) to reinforce the high-quality response in reasoning tasks (Lightman et al., 2024). These approaches reduce labeling cost by using surrogate supervision to guide the optimization, which drives the model toward a higher level of intelligence automatically. However, current self-reward signals easily induce training collapse (Shafayat et al., 2025) as shown in Figure 1, and thus are limited to scale up on larger foundation models.

This requires us to rethink the underlying belief of realizing self-evolving without external feedback (Wei et al., 2022; Shafayat et al., 2025; Zhao et al., 2025b), or more fundamentally, how to characterize latent reasoning capability beyond the empirical passing rate to specific questions. Conceptually, it can be understood as a kind of abstract knowledge that consists of implicit logical information (e.g., elementary arithmetic). Although confident outputs are prone to being the correct answer, it does not guarantee that the model is based on a valid reasoning path. Instead of solely relying on output-side signals representing necessary results, we should also complement the eliciting process with validity confirmation, which is more aligned with the induction paradigm for the symbolic mechanism (Yang et al., 2025b). Naturally, it brings us back to conventional self-supervised learning (Tsai et al.; Liu et al., 2021) that captures intrinsic representation via input-level variants (Hendrycks & Dietterich), which initiates a key hypothesis, i.e. *contrastive agreement*, for our intuition on reasoning refinement.

In this work, we introduce *Co-Reward*, a new self-supervised reinforcement learning framework designed to elicit reasoning capability through contrastive agreement. In the high level, Co-Reward constructs structured training signals from input-side question variants (refer to Figure 2), and encourages model to capable with valid reasoning ability. Specifically, for a given input, we generate a set of rephrased prompts with equivalent semantic intent. The model’s responses to these analogs are aggregated by majority voting to form a pseudo-consensus answer, which is then used to supervise the original input through a reward signal that reflects the alignment with the dominant reasoning trajectory. Based on the above atom operation, Co-Reward forms a dual-path structure to shaping the reward: each majority-voted answers serve as cross-reference to the paired questions for optimization, which explicitly expand the input coverage for reasoning validity.

Using the contrastive agreement as a proxy, Co-Reward encourages the model to reinforce reasoning strategies that are robust to linguistic variation, while suppressing trivial solutions that induce model collapse for fake reward. It provides a principled self-training signal that requires no ground-truth labels, yet elicit an intrinsic reasoning ability that valid for broader input coverage. Empirically, we evaluate Co-Reward on a series of reasoning benchmarks using different base models such as the Llama (Dubey et al., 2024) and Qwen (Qwen et al., 2025) series. It outperforms other self-reward baselines across multiple reasoning benchmarks and LLM series, and achieves performance comparable to or exceeding that of GT labeled reward, with improvements of up to +6.8% over GT reward on Llama-3.2-3B-Instruct (refer to Table 1). In addition, it can also achieve the superior *test-time training* (refer to Table 2) performance due to its independence on ground-truth label. We observe that Co-Reward leads to more stable optimization and mitigates collapse behaviors, and conduct ablation studies to provide a comprehensive understanding. These findings demonstrate the promise of Co-Reward for unleashing the latent reasoning capability. To summarize, our main contributions are as follows,

- We introduce a new perspective, from self-supervised learning, to elicit latent reasoning capability via contrastive agreement, which aligns with the induction paradigm from input-side information.
- We propose Co-Reward, a novel self-supervised reinforcement learning framework that forms a dual-path to construct self-generate rewards and cross-overly supervise input variants.
- We empirically demonstrate the general effectiveness of Co-Reward to achieve superior RL and test-time training performance, and also present various ablation studies and further analyses.

2 PRELIMINARY

Problem Setups. Given a large language model π_θ parameterized by θ , and a dataset D consisting of question-answer pairs (x, a) , the model generates a response $y \sim \pi_\theta(\cdot|x)$ auto-regressively. In particular, $y = (y^1, y^2, \dots, y^n)$ and each token is generated according to the conditional probability $y^{k+1} \sim \pi_\theta(\cdot|x, y^{<k})$ given $y^{<k}$ denotes the first generated k tokens. We consider the LLM output with the step-by-step reasoning process and the predicted answer to the question, and a verifiable reward function $r(x, y)$ that can extract the answer with $\text{ans}(\cdot)$ and evaluate the output correctness with the prompt-specific ground-truth answer a as,

$$r(a, y) = \begin{cases} 1 & \text{If ans}(y) \text{ is correct with answer } a, \\ 0 & \text{If ans}(y) \text{ is incorrect with answer } a. \end{cases} \quad (1)$$

Then the general objective of training LLM for complex reasoning via RLVR (Shao et al., 2024; Yu et al., 2025) can be formulated with the policy model π_θ as,

$$\max_{\pi_\theta} \mathbb{E}_{y \sim \pi_\theta(x)} [r(a, y) - \beta \cdot \text{KL}[\pi_\theta(y|x) || \pi_{\text{ref}}(y|x)]], \quad (2)$$

where π_{ref} is an initial reference policy, and β is a coefficient controlling the KL divergence to prevent excessive deviation from the reference model. Intuitively, the training target is to maximize the reward in passing specific reasoning questions while maintaining the general capability of LLM.

In practice, various methods are developed in RLVR based on the classic REINFORCE (Williams, 1992) algorithm, such as REINFORCE++ (Hu, 2025), PPO (Schulman et al., 2017), and GRPO (Shao et al., 2024). To narrow down our scope, we introduce the representative and widely adopted Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Different from PPO adopting the learnable value model. Instead, GRPO utilizes the average reward of multiple sampled outputs for the same question. Specifically, given a question x , GRPO requires to sample G outputs from the old policy π_{old} as $\{y_i\}_{i=1}^G \sim \pi_{\text{old}}(\cdot|x)$. Then, it computes the reward r_i for each output y_i (through deterministic reward functions) and obtains a group of rewards $\{r_i(y_i)\}_{i=1}^G$. The advantage \hat{A}_i is estimated as:

$$\hat{A}_i = \bar{r}_i = \frac{r_i(a_i, y_i) - \text{mean}(\{r_i(a_i, y_i)\}_{i=1}^G)}{\text{std}(\{r_i(a_i, y_i)\}_{i=1}^G)}. \quad (3)$$

Then the target policy is optimized by maximizing the advantage while ensuring the policy model remains close to the reference policy:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim D, \{y_i\}_{i=1}^G \sim \pi_{\text{old}}(Y|x)} \underbrace{\frac{1}{G} \sum_{i=1}^G \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left(\min \left[c_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(c_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \right] - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \right)}_{\mathcal{R}(\hat{A})}, \quad (4)$$

where

$$c_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t}|x, y_{i,<t})}, \quad \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\text{ref}}(y_{i,t}|x, y_{i,<t})} - \log \frac{\pi_{\text{ref}}(y_{i,t}|x, y_{i,<t})}{\pi_\theta(y_{i,t}|x, y_{i,<t})} - 1. \quad (5)$$

Note that the $\text{clip}(\cdot, 1 - \epsilon, 1 + \epsilon)$ in Eq. (4) is used to ensure that updates do not deviate excessively from the old policy by bounding the policy ratio between $1 - \epsilon$ and $1 + \epsilon$ in a risk function $\mathcal{R}(\hat{A})$.

Self-reward RL. Without relying on the ground-truth answer like conventional GRPO in various reasoning tasks, recent works (Agarwal et al., 2025; Shafayat et al., 2025; Zhao et al., 2025b) also explore self-reward training for learning to reason without external reward. These methods employ self-generated training signals to supervise the learning process. In general, they can be summarized into two categories: *entropy*-based or *consistency*-based. The former directly minimize the Shannon entropy $\mathcal{H}(\pi_\theta) = -\mathbb{E}_{y \sim \pi_\theta(x)} [\log \pi_\theta(x)]$ (Shannon, 1948) of the policy model π_θ , either estimated by trajectory distribution (Tiapkin et al., 2023) or token distribution (Haarnoja et al., 2018) at each step. The latter propose to utilize evolving majority votes (Wang et al., 2022) as a pseudo-label of the ground truth. Specifically, the training iteration of the pipeline (Shafayat et al., 2025) is processed as generating n solutions $y_1, y_2, \dots, y_n \sim \pi_\theta(\cdot|x)$ for each question, identifying the majority-voted

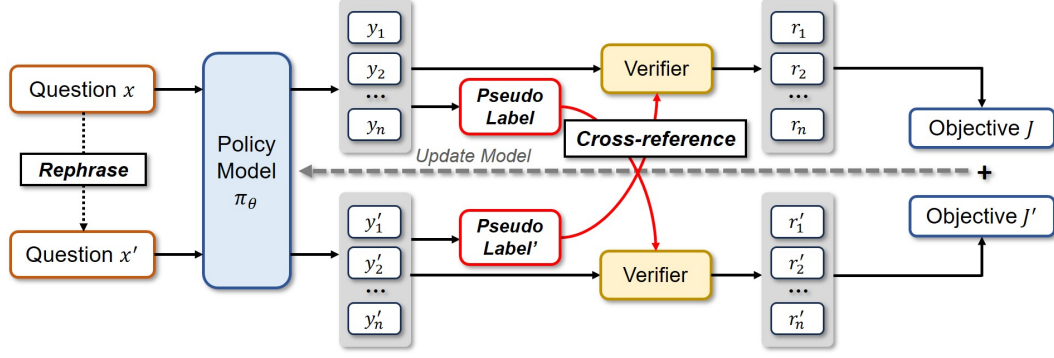


Figure 2: Illustration of *Co-Reward* Framework: given the original questions in training set, we generate the paraphrased version for model to response, then we use the majority-voted answers of each as pseudo-labels to cross-overly supervise the model learning on original and new questions.

answer $y_v \leftarrow \arg \max_{y^*} \sum_{i=1}^n 1[\text{ans}(y_i) = \text{ans}(y^*)]$, and then use the agreement with voted answer as the binary reward $r(y_v, y) = 1[\text{ans}(y) = \text{ans}(y_v)]$ for policy optimization. Intuitively, these approaches share the similar intuition on boosting model confidence as an proxy target for reasoning tasks, and are demonstrated to be promising on improving the performance on Qwen series models.

Nonetheless, self-reward RL is unstable during training, especially on the later stage, it is easy to induce learning collapse where the model consistently generate wrong output and show extreme confidence on the answer, which achieve the proxy target without genuinely enhancing reasoning.

3 CO-REWARD: SELF-SUPERVISED ANALOGY

Given the high cost of human annotation and the growing complexity of reasoning tasks (Song et al., 2024; Ouyang et al., 2022), an emerging and promising direction for scalable post-training of LLMs is to enhance reasoning capabilities without relying on ground-truth labels (Shafayat et al., 2025; Zhao et al., 2025b), which relies solely on unlabeled data and aligns with the vision of self-evolving.

In the following, we present Co-Reward, a novel self-supervised reinforcement learning framework for LLM to elicit the latent reasoning capability through the intuition of contrastive agreement.

3.1 ELICITING LATENT REASONING CAPABILITY OF LLMs

One prerequisite of self-evolving is that LLMs possess substantial latent reasoning capabilities and can be post-refined without external supervision. This is firstly grounded in the observation (Song et al., 2024; Zelikman et al., 2022) that pretrained LLMs already exhibit non-trivial reasoning behaviors across a wide range of tasks, despite never being explicitly trained to do so. The behaviors emerge from pretraining on vast corpora (Achiam et al., 2023) that implicitly encode logical structures (Wei et al., 2022), analogous patterns (Yuan et al., 2024), and procedural knowledge (Kosinski, 2024).

Previous self-reward methods (Shafayat et al., 2025; Zhao et al., 2025b) mainly focus on utilizing the output-side information to construct proxy target, like encouraging or selecting confident model output, which are prone to be high-quality and approach to the ground truth. The underlying belief is that prompting model self-certainty would induce the reasoning power in high probability. Although empirically effective, the paradigm lacks reasoning validity confirmation that ensure the reliability of the encouraged behaviors. Recalling the conceptual definition of latent reasoning capability, it should be possible to directly originate from data instead of solely defined on consistency among repeated model behaviors. Which is to say, the fundamental induction can be more model-independent.

Based on the previous conjecture, we take inspiration from the broader field of *self-supervised learning* (Bengio et al., 2013; Tsai et al.; Chen et al., 2020; Khosla et al., 2020; Liu et al., 2021), where models learn useful features by solving auxiliary tasks defined purely over unlabeled data. Similarly, we propose to view reasoning as a structured function that should remain stable under semantically

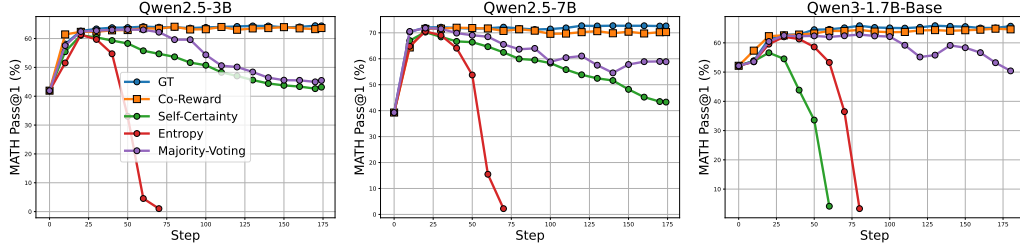


Figure 3: Validation curves on MATH5000 of Co-Reward with baselines. We can find the learning of Co-Reward is generally stable without significant performance drop across different LLMs.

equivalent transformations. If the model’s outputs across such variants exhibit convergence, then this agreement can be treated as a reliable signal for containing the reasoning path.

3.2 CONTRASTIVE AGREEMENT AS IMPLICIT INDUCTION

Given the intrinsic knowledge contained in LLM, using RL for reasoning refinement can be regarded as reorganization and utilization for further infer the valid results, which needs induction on the specific questions from the previous view of understanding. Thus, the previous conceptual lens motivates us to considering the input-side transformation beyond identifying confident model behavior.

We hypothesize one key inductive property of LLMs corresponding to the possibility of eliciting valid reasoning capability: the outputs tend to be intrinsically self-consistent across semantically equivalent inputs. That is, if a reasoning chain is valid, then small transformation in the rephrasing of the question context (e.g., via paraphrasing, reordering, or reformatting) should still lead the model to recover the same or similar answer. This forms the foundation for a self-referential training signal: *contrastive agreement* among different question variants can serve as an optimization proxy.

While regularizing the latent space like traditional contrastive learning is hard in RL framework, we define contrastive agreement as a principle that aligns model reasoning outputs, treating consistent inter-view agreement as a signal for valid inference. This complements single-view self-reward strategies by introducing a form of collective validity verification with broader input consideration.

It states that consistent outputs across semantically equivalent inputs reflect an implicit alignment with an underlying reasoning structure. Rather than relying on external correctness signals or confidence-based conditions, contrastive agreement exploits the model behavior to reinforce stable reasoning paths. It presumes that when a model consistently arrives at the same conclusion through differently rephrased questions, the associated reasoning trajectory is more likely to be valid and generalizable.

3.3 ALGORITHMIC REALIZATION

Building upon the discussed contrastive agreement, we present our Co-Reward framework as illustrated in Figure 2. Formally, its learning objective can be formulated as follows based on GRPO,

$$\mathcal{J}_{\text{Co-Reward}}(\theta) = \underbrace{\mathbb{E}_{x \sim D, \{y_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(Y|x)} \mathcal{R}(\hat{A})}_{\mathcal{J}_{\text{original}}(\theta)} + \underbrace{\mathbb{E}_{x' \sim D', \{y'_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(Y'|x')} \mathcal{R}(\hat{A}')}_{\mathcal{J}_{\text{rephrased}}(\theta)}, \quad (6)$$

where the relative advantages are estimated by the *cross-refereed* supervision as,

$$\hat{A}_i = \frac{r_i(y'_v, y_i) - \text{mean}(\{r_i(y'_v, y_i)\}_{i=1}^G)}{\text{std}(\{r_i(y'_v, y_i)\}_{i=1}^G)}, \quad \hat{A}'_i = \frac{r_i(y_v, y'_i) - \text{mean}(\{r_i(y_v, y'_i)\}_{i=1}^G)}{\text{std}(\{r_i(y_v, y'_i)\}_{i=1}^G)}. \quad (7)$$

Specifically, given a set of original questions, we utilize the rephrased version that keeps the semantical equivalence for the model to respond, and then collect the self-generated pseudo-labels based on the majority voting mechanism (Shafayat et al., 2025) as follows to supervise learning on the counterparts,

$$y_v \leftarrow \arg \max_{y^*} \sum_{i=1}^n 1[\text{ans}(y_i) = \text{ans}(y^*)], \quad y'_v \leftarrow \arg \max_{y^*} \sum_{i=1}^n 1[\text{ans}(y'_i) = \text{ans}(y^*)]. \quad (8)$$

The overall pipeline can be viewed as a dual-path structure with cross-reference in the reward shaping process, it may also be compatible with other self-generated feedbacks (Wang et al., 2022) on the output-side information due to the generality of the core idea. While in the current version, we choose the majority voting mechanism in the implementation for the empirical effectiveness and simplicity.

We summarize the whole realization of this framework in Algorithm 1. Unlike conventional contrastive learning that aligns representation vectors, our contrastive objective operates on self-generated reasoning answers, encouraging the model to align its reasoning results to different questions that share semantic intent. Formally, for each input prompt, the Co-Reward signal increases when the model’s output is consistent with the majority answer ob-

Algorithm 1 Co-Reward

Input: policy model: π_θ , training learning rate: η , training dataset: \mathcal{D} , rephrased training dataset: \mathcal{D}' , total iteration: T ;

Output: model θ_T ;

```

1: for all iteration do
2:   Sample mini-batch inputs from  $\mathcal{B} \subseteq \mathcal{D}$  and  $\mathcal{B}' \subseteq \mathcal{D}'$ 
3:   for all input prompt  $x \in \mathcal{B}$  and  $x' \in \mathcal{B}'$  do
4:     Generate  $n$ -paired responses on  $y_1, y_2, \dots, y_n \sim \pi_\theta(\cdot|x)$  and
        $y'_1, y'_2, \dots, y'_n \sim \pi_\theta(\cdot|x')$ 
5:     Identify majority-voted answers by Eq. (8)
6:     Estimate the relative advantages by Eq. (7)
7:     Calculate the policy objective by Eq. (6)
8:     Update parameters:  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{J}_{\text{Co-Reward}}(\theta)$ 
9:   end for
10: end for
```

tained from its analogical counterparts, and decreases when it diverges. This contrastive agreement promotes semantic invariance, implicitly increasing the difficulty of reaching trivial solutions to obtain the reward (e.g., achieving the arbitrary answers but consistent on single input), which can be found in experimental part verification.

By shaping the reward around inter-view agreement, Co-Reward enables scalable and self-supervised reasoning refinement, circumventing the need for human-labeled correctness. Intuitively, our framework enhance model’s robustness to linguistic variations and also empirically alleviate the model collapse, as demonstrated in Figure 3, with a broader consideration on different inputs.

4 EXPERIMENTS

4.1 SETUPS

Backbone Models and Baselines. We utilize a diverse set of LLMs with varying architectures and scales, which includes the Qwen2.5 series (Qwen et al., 2025) (Qwen2.5-3B/7B), the Qwen3 series (Yang et al., 2025a) (Qwen3-1.7B/4B/8B-Base) and the Llama3 series (Dubey et al., 2024) (Llama3.2-3B-Instruct). Beyond the vanilla GRPO that utilizes GT for rewards (Shao et al., 2024), we benchmark Co-Reward against several state-of-the-art approaches, including Self-Certainty (Zhao et al., 2025b), Entropy (Prabhudesai et al., 2025) and Majority Voting (Shafayat et al., 2025).

Implementation Details. We implement our algorithm on the top of the VeRL framework (Sheng et al., 2024). Experiments are conducted on $4 \times$ NVIDIA H200 GPUs. Specifically, all of the methods are trained on the training split of the MATH dataset (Hendrycks et al., 2021), which contains 7,500 problems. To ensure a fair comparison, we utilize the officially released chat-based prompting formats for all experiments. For every RL update during training, we sample 128 problems and generate 8 candidate rollouts per problem, with a learning rate of $3e-6$ and KL penalty of $\beta = 0.005$. Additional hyperparameter setups can be found in Appendix C.

Evaluation Datasets and Metrics. To provide a comprehensive evaluation of model capabilities, we utilize a diverse set of benchmarks spanning mathematical reasoning, code generation, instruction-following, and general multi-task abilities. Specifically: (1) Mathematical reasoning: We evaluate on MATH500 (Lightman et al., 2024), GSM8K (Cobbe et al., 2021), and AMC (Li et al., 2024a). For MATH500 and GSM8K, we report pass@1 accuracy using the `lighteval` library¹. For AMC, we use the `ttrl`² library and report avg@8 as the metric. (2) Code generation: We assess coding

¹<https://github.com/huggingface/lighteval>

²https://github.com/ruixin31/Spurious_Rewards/tree/main/code/ttrl

Table 1: Main Results (%) of *RL performance* comparison on reasoning benchmarks. Cell background colors indicate relative performance: darker colors denote better results within each model group.

Methods	Mathematics			Code		Instruction	Multi-Task
	MATH500	GSM8K	AMC	LiveCode	CRUX	IFEval	MMLU-Pro
Qwen2.5-3B							
Before RL	53.6	19.48	10.69	9.95	18.50	29.83	32.50
- GT-Reward (Shao et al., 2024)	65.4	82.18	32.98	13.93	32.12	33.66	36.74
- Self-Certainty (Zhao et al., 2025b)	64.2	80.52	28.92	10.90	29.00	32.22	33.88
- Entropy (Prabhudesai et al., 2025)	63.2	80.44	29.67	9.05	29.00	32.94	35.35
- Majority-Voting (Shafayat et al., 2025)	64.6	82.41	33.13	14.03	36.38	35.19	35.50
- Co-Reward (Ours)	65.4	84.53	30.57	16.40	36.88	33.86	36.38
Qwen2.5-7B							
Before RL	69.4	24.71	15.81	3.79	26.38	38.19	44.76
- GT-Reward (Shao et al., 2024)	76.4	88.02	45.63	15.92	45.12	41.49	41.12
- Self-Certainty (Zhao et al., 2025b)	72.8	84.31	38.55	12.04	54.12	37.24	43.30
- Entropy (Prabhudesai et al., 2025)	72.2	81.43	39.61	16.49	51.88	40.33	42.79
- Majority-Voting (Shafayat et al., 2025)	74.4	84.53	40.96	15.45	51.00	38.60	43.35
- Co-Reward (Ours)	74.6	89.61	41.27	15.73	55.58	42.86	40.51
Qwen3-1.7B-Base							
Before RL	57.0	19.56	8.43	4.45	7.50	33.65	33.00
- GT-Reward (Shao et al., 2024)	69.6	81.57	35.54	13.74	35.25	36.16	39.12
- Self-Certainty (Zhao et al., 2025b)	58.2	40.25	23.04	9.86	18.00	32.96	35.13
- Entropy (Prabhudesai et al., 2025)	63.6	71.79	31.63	13.74	31.37	35.37	36.67
- Majority-Voting (Shafayat et al., 2025)	65.2	81.57	34.78	13.08	34.25	35.45	36.00
- Co-Reward (Ours)	67.6	83.01	32.22	13.50	32.38	35.56	35.53
Qwen3-4B-Base							
Before RL	71.2	26.15	21.08	11.00	38.88	46.43	47.23
- GT-Reward (Shao et al., 2024)	78.6	89.76	51.20	26.07	55.38	47.80	53.96
- Self-Certainty (Zhao et al., 2025b)	71.6	71.79	38.86	22.37	57.00	48.15	48.93
- Entropy (Prabhudesai et al., 2025)	77.0	88.10	47.44	25.59	52.88	50.44	49.90
- Majority-Voting (Shafayat et al., 2025)	77.4	90.07	45.33	26.54	57.50	48.78	54.35
- Co-Reward (Ours)	78.8	91.28	46.08	26.64	56.50	50.35	53.26
Qwen3-8B-Base							
Before RL	72.4	27.82	20.93	23.41	54.75	50.89	52.92
- GT-Reward (Shao et al., 2024)	82.6	87.26	54.22	30.52	63.25	52.78	57.11
- Self-Certainty (Zhao et al., 2025b)	80.2	80.74	50.75	27.20	64.38	50.98	54.17
- Entropy (Prabhudesai et al., 2025)	80.2	87.19	49.54	29.38	62.00	51.81	54.86
- Majority-Voting (Shafayat et al., 2025)	79.8	89.76	49.09	30.52	63.38	51.80	56.93
- Co-Reward (Ours)	81.2	93.70	51.20	30.81	66.00	55.79	59.95
Llama-3.2-3B-Instruct							
Before RL	39.2	65.73	10.54	9.86	25.37	57.32	31.14
- GT-Reward (Shao et al., 2024)	47.0	77.94	22.14	9.57	31.87	47.51	34.32
- Self-Certainty (Zhao et al., 2025b)	43.4	74.91	18.83	9.95	25.87	54.88	33.34
- Entropy (Prabhudesai et al., 2025)	43.4	66.19	20.18	11.66	24.62	54.70	33.52
- Majority-Voting (Shafayat et al., 2025)	46.8	78.77	20.48	11.00	31.25	47.96	33.18
- Co-Reward (Ours)	50.2	79.45	23.80	11.28	29.88	48.89	33.77

ability using LiveCodeBench (Jain et al., 2025) release_v6 and CRUX. LiveCodeBench is evaluated with its official evaluation library³, and CRUX (Gu et al., 2024) is evaluated via the ZeroEval library⁴; for both datasets, we report pass@1 accuracy. (3) Instruction-following and multi-task abilities: We evaluate on IFEval (Zhou et al., 2023b) and MMLU-Pro (Wang et al., 2024), using the lm-evaluation-harness library⁵ for both. Additional details are provided in Appendix D.

4.2 RL PERFORMANCE COMPARISON

Co-Reward achieves the Competitive performance among self-reward baselines. As shown in Table 1, we evaluate Co-Reward against multiple self-reward baselines on various downstream

³<https://github.com/LiveCodeBench/LiveCodeBench>

⁴<https://github.com/WildEval/ZeroEval>

⁵<https://github.com/EleutherAI/lm-evaluation-harness>

Table 2: Main Results (%) of *test-time training* for boosting reasoning performance.

LLMs	Methods	AMC			
		avg@8	avg@16	avg@32	avg@64
<i>Qwen2.5-7B</i>	Before-TTT	15.81	17.55	16.34	17.32
	Self-Certainty	22.89	23.64	21.91	21.57
	Entropy	21.84	24.17	21.88	22.29
	Majority-Voting	17.92	21.69	20.67	19.86
	Co-Reward (Ours)	25.45	23.64	25.08	24.34
<i>Qwen3-8B-Base</i>	Before-TTT	20.93	21.31	19.58	20.97
	Self-Certainty	40.66	40.06	39.00	38.50
	Entropy	36.45	32.83	32.53	33.30
	Majority-Voting	38.25	36.97	35.80	36.42
	Co-Reward (Ours)	44.28	43.60	42.58	43.43

tasks. Co-Reward consistently achieves state-of-the-art performance, demonstrating its superior effectiveness. Intuitively, Co-Reward occupies more darker areas in this Table. Specifically, on the three mathematics benchmarks, Co-Reward achieves an average relative performance gain of 2.35% with Qwen2.5-7B and 2.18% with Qwen3-8B-Base. Remarkably, on Llama-3.2-3B-Instruct, Co-Reward yields an average relative gain of 8.11%, highlighting its strong generalization capability across diverse LLM architectures. Moreover, Co-Reward unlocks considerable potential in the code domain, resulting in substantial improvements on related downstream tasks. Importantly, Co-Reward training does not compromise the instruction-following and multi-task abilities of the base models, and even leads to improvements on certain models.

Co-Reward enables further refinement of LLMs during inference. Self-supervised methods, which do not require ground-truth labels, are naturally compatible with test-time training (TTT). This allows models to further refine themselves during inference. We conduct test-time training experiments on the challenging competition-level dataset AMC. Table 2 summarizes the experimental results. It can be observed that our Co-Reward method achieves consistently strong performance across all settings. In particular, on Qwen3-8B-Base, Co-Reward outperforms all baselines by a clear margin for all values of k . These results demonstrate the effectiveness of Co-Reward in enhancing the reasoning ability of LLMs during inference, especially when applied to more powerful models such as Qwen3-8B-Base. Moreover, the higher avg@ k scores reflects that Co-Reward enhances the robustness and stability of the LLM in generating correct answers across multiple rollouts.

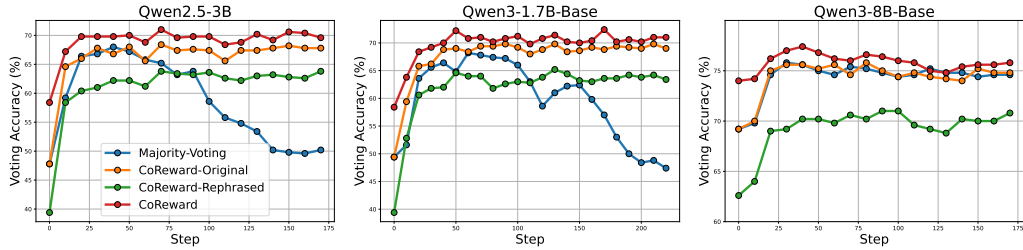


Figure 4: Voting accuracy (%) across different LLMs with increasing training steps. **Majority-Voting** measures the voting accuracy during the “Majority-Voting” training. **CoReward-Original** and **CoReward-Rephrased** record the voting accuracies during Co-Reward training, on the original and rephrased math problems respectively. **CoReward** counts the voting accuracy if either the original or rephrased voted answer is correct, reflecting the pseudo-label cross-reference strategy.

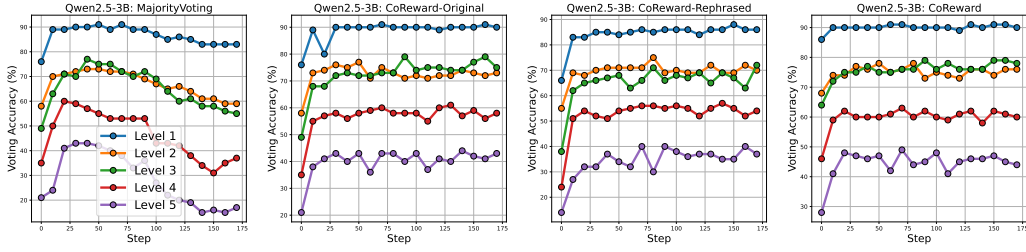


Figure 5: Voting accuracy across different difficulty level of math problems.

Table 3: The impact of rephrased math problems.

Methods	MATH500	GSM8K	AMC	LiveCode	CRUX	IFEval
<i>Qwen3-8B-Base</i>						
Majority-Voting w/ Original	79.8	89.76	49.09	30.52	63.38	51.80
Majority-Voting w/ Rephrased	79.2	91.51	50.75	31.66	60.38	52.24
Co-Reward	81.2	93.70	51.20	30.81	66.00	59.95

4.3 ANALYSIS OF VOTING ACCURACY DURING TRAINING

To better understand the effectiveness of the proposed Co-Reward approach, we conduct a statistical analysis of the voting accuracy during the training process. The MATH dataset contains 7,500 training math problems across five difficulty levels. We randomly select 100 problems from each difficulty level to form a balanced subset of 500 problems for the analysis. From the perspectives of methodology and problem difficulty, we have two core observations respectively:

Co-Reward produces more stable and robust voted answers. Figure 4 illustrates the voting accuracy across three LLMs as training progresses. It can be observed that the Co-Reward, along with its one-side variants CoReward-Original and CoReward-Rephrased consistently maintains stable voting accuracy compared to the Majority-Voting baseline. The observed stability can be largely attributed to the cross-reference pseudo-labeling strategy, where original and rephrased problems mutually supervise each other. This mutual supervision effectively mitigates potential reward hacking caused by a single-view consistency bias, thus preventing training instability and collapse.

Co-Reward effectively mitigates the collapse of voting strategy on challenging problems. Figure 5 illustrates the voting accuracy for different-level math problems (Level 1-5), with Level 1 being the easiest and Level 5 the most difficult. It can be observed that Majority-Voting suffers a noticeable drop in accuracy on more difficult problems, especially at Levels 4 and 5. In contrast, Co-Reward and its variants consistently maintain stable accuracy across all difficulty levels. This robustness suggests that Co-Reward is better equipped to handle challenging problems during training, which is a key factor underlying its overall superior performance compared to other self-supervised methods. The

4.4 FURTHER ANALYSIS

The impact of the rephrased data. To investigate the impact of the rephrased math problems, we train Majority-Voting directly on either the original or rephrased problems using Qwen3-8B-Base. Table 3 summarizes the experimental results, where “w/ Original” denotes training on original math problems, and “w/ Rephrased” refers to training on rephrased problems. It can be observed that, across all benchmarks, training with rephrased problems generally yields performance comparable to training with the original problems, indicating the high quality of the rephrased data. Notably, our Co-Reward approach, which jointly leverages both original and rephrased data through cross-referenced pseudo-labeling, consistently achieves superior results on all benchmarks except LiveCode. This highlights that mutual supervision from two different perspectives is both reasonable and effective for improving overall model performance.

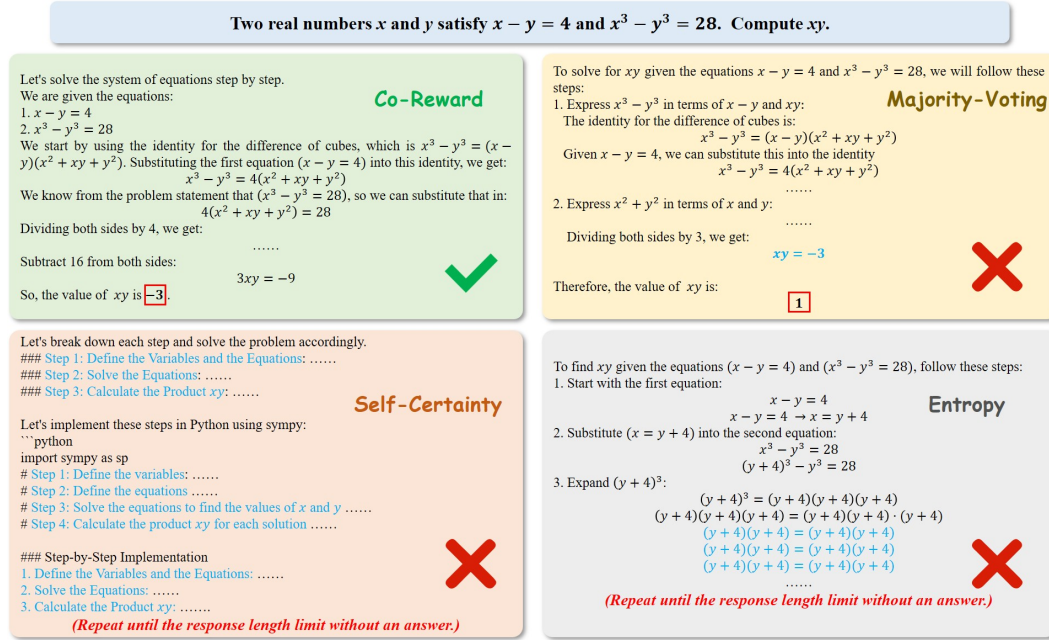


Figure 6: Case study of the Qwen3-1.7B-Base generation trained by four self-supervised methods.

Case study of generations across different methods. To intuitively demonstrate the qualitative differences among different methods, we present a case study comparing the generated content for the same math problem across Co-Reward, Majority-Voting, Self-Certainty and Entropy. As illustrated in Figure 6, our Co-Reward method produces a well-structured reasoning process and correctly arrives at the final answer. Interestingly, although Majority-Voting also outputs an answer, it erroneously places “1” into the answer box. This occurs because Majority-Voting can be easily hacked: when n rollouts consistently output an answer such as “1” in the box, the method attains highest reward regardless of the correctness of the underlying reasoning. Such reward hacking ultimately leads to training collapse. Furthermore, both Self-Certainty and Entropy suffer from the repetition problem, failing to produce a complete answer. This is because optimizing purely for high self-certainty or low entropy encourages the model to concentrate probability mass on a small set of tokens, often resulting in repetitive outputs. In such cases, the model achieves high self-certainty or low entropy simply by repeating the same tokens, resulting in model collapse. The complete generated outputs for this case are provided in Appendix D.4.

5 RELATED WORK

Large Language Model Reasoning. LLMs have shown impressive performance on vast tasks that require reasoning, including solving mathematical problems, writing code, and answering logical questions. One of the key techniques that has improved LLM reasoning is Chain-of-Thought (CoT) prompting (Wei et al., 2022). CoT encourages the model to generate intermediate reasoning steps before producing the final answer, which has been shown to enhance performance on tasks like arithmetic, commonsense reasoning, and symbolic reasoning. Subsequent work has extended CoT by integrating it with various strategies, including compositional generalization (Zhou et al., 2023a; Khot et al., 2023) and employing structural reasoning approaches (Yao et al., 2023a; Besta et al., 2024; Yang et al., 2024). In addition, CoT serves as a fundamental framework for techniques like fine-tuning (Zelikman et al., 2022), argentic workflow (Yao et al., 2023b), and paving the way for inference-time scaling (Snell et al., 2024).

RL for Large Language Models. Several RL algorithms have been developed primarily for alignment tasks. Specifically, DPO (Rafailov et al., 2023), CPO (Xu et al., 2024), and their variants (Li

et al., 2024b; Guo et al., 2024; Munos et al., 2024; Hong et al., 2024; Xie et al., 2024) rely on pairs of outputs labeled by human preference. In contrast, KTO (Ethayarajh et al., 2024) and BCO (Jung et al., 2024) require only a single binary label (like or dislike) for each output. Besides, the PRM (Uesato et al., 2022; Lightman et al., 2024) and Step-KTO (Lin et al., 2025a) offer step-by-step guidance by incorporating feedback at each reasoning step rather than focusing solely on the final outputs. Recently, the follow-up work of GRPO improves the optimization objective, *e.g.*, DAPO (Yu et al., 2025), Dr. GRPO (Liu et al., 2025a), REINFORCE++ (Hu, 2025), CPPO (Lin et al., 2025b), and GPG (Chu et al., 2025). Another line of research generalizes GRPO to broader applications such as multimodal reasoning (Zhou et al., 2025; Huang et al., 2025; Chu et al., 2025; Liu et al., 2025b; Zhang et al., 2025a) and logical reasoning (Xie et al., 2025).

RL without External Reward. RL methods have shown promising scaling capabilities to enhance the reasoning abilities of LLMs (Guo et al., 2025), yet they are often limited by the availability of training data for reward signals (Gao et al., 2023; Liu et al., 2023). Notably, Wang et al. (2025) demonstrates that RL can effectively bootstrap LLM reasoning with as little as a single training example, highlighting the potential to minimize or even eliminate reliance on external reward signals during training. Recent efforts leverage distinct strategies for reward assignment. For instance, SIRLC (Pang et al., 2024) and AZR (Zhao et al., 2025a) utilize an LLM-as-the-judge approach to assign rewards. In contrast, methods like SRT, TTRL, and their variants (Shafayat et al., 2025; Zuo et al., 2025; Fang et al., 2025; Zhang et al., 2025b) employ self-consistency (Wang et al., 2022) to generate pseudo-rewards, reducing dependence on external annotations. Meanwhile, INTUITOR, RLSC, and RENT (Zhao et al., 2025b; Li et al., 2025; Prabhudesai et al., 2025) harness the internal confidence scores of LLMs as intrinsic reward signals. Additionally, EMPO and its variants (Zhang et al., 2025c; Agarwal et al., 2025) promote reasoning by minimizing entropy during the reasoning process, further diversifying the approaches to incentivize robust LLM performance.

6 CONCLUSION

In this work, we proposed Co-Reward, a self-supervised reinforcement learning framework designed to elicit the latent reasoning capabilities of LLM via contrastive agreement. By leveraging analogy across semantically equivalent question rewrites and employing majority voting, Co-Reward constructs robust, cross-referable reward signals without requiring explicit labels, aligning reinforcement learning with reasoning consistency rather than single-instance outputs prone to learning collapse. This interview agreement-based reward shaping considers broader input-side coverage while promoting stable and genuine reasoning improvements. We hope this work will inspire further exploration into self-supervised RL for reasoning to advance the development of LLM-based reasoning systems.

REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI*, 2024.
- Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.
- Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.
- Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. Serl: Self-play reinforcement learning for large language models with limited data. *arXiv preprint arXiv:2505.20347*, 2025.
- Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.
- Alex Gu, Baptiste Roziere, Hugh James Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida Wang. Cruxeval: A benchmark for code reasoning, understanding and execution. In *International Conference on Machine Learning*, pp. 16568–16621. PMLR, 2024.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.
- Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

- Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021.
- Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.
- Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-rl: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*, 2023.
- Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121, 2024.
- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024a.
- Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence is all you need: Few-shot rl fine-tuning of language models. *arXiv preprint arXiv:2506.06395*, 2025.
- Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *ICML*, 2024b.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *ICLR*, 2024.
- Yen-Ting Lin, Di Jin, Tengyu Xu, Tianhao Wu, Sainbayar Sukhbaatar, Chen Zhu, Yun He, Yun-Nung Chen, Jason Weston, Yuandong Tian, et al. Step-kto: Optimizing mathematical reasoning through stepwise binary feedback. *arXiv preprint arXiv:2501.10799*, 2025a.
- Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 2025b.

- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36:21558–21572, 2023.
- Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding rl-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025a.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025b.
- Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level, 2025.
- Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. In *ICML*, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. Language model self-improvement by reinforcement learning contemplation. In *The Twelfth International Conference on Learning Representations*, 2024.
- Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train? *arXiv preprint arXiv:2505.21444*, 2025.
- Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.
- Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. *arXiv preprint arXiv:2412.02674*, 2024.

- Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, pp. 34161–34221. PMLR, 2023.
- Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.
- Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.
- Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.
- Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.
- Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. Buffer of thoughts: Thought-augmented reasoning with large language models. In *NeurIPS*, 2024.
- Yukang Yang, Declan Iain Campbell, Kaixuan Huang, Mengdi Wang, Jonathan D. Cohen, and Taylor Whittington Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Forty-second International Conference on Machine Learning*, 2025b. URL <https://openreview.net/forum?id=y1SnRPDwX4>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023a.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023b.

- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.
- Siyu Yuan, Jiangjie Chen, Changzhi Sun, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. Analogymb: Unlocking analogical reasoning of language models with a million-scale knowledge base. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1249–1265, 2024.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025a.
- Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08745*, 2025b.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *arXiv preprint arXiv:2504.05812*, 2025c.
- Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025a.
- Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025b.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023a.
- Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023b.
- Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, et al. Ttrl: Test-time reinforcement learning. *arXiv preprint arXiv:2504.16084*, 2025.

APPENDIX

A BROADER IMPACT

This paper focuses on learning RLVR without without reliance on manually annotated ground truth data. This issue is especially critical when leveraging vast, unlabeled datasets or operating under budgetary constraints. In contrast to prevailing methods, our proposed framework, Co-Reward, is built on a self-supervised learning paradigm. It introduces a novel consistency-based contrastive reward signal across semantically analogical questions, that promotes stable reasoning elicitation and improvement. This research direction holds great promise for further enhancement through more detailed and refined studies.

B LIMITATION AND FUTURE WORKS

While the proposed Co-Reward method demonstrates promising results, there are several limitations that warrant attention. A primary limitation is the introduction of additional computational overhead associated with generating rollouts for augmented questions. Furthermore, although Co-Reward enhances training stability, the risk of reward hacking persists, particularly when question prompts are excessively complex and fall outside the scope of the model’s pre-trained knowledge. Future work should therefore focus on developing more computationally efficient strategies and exploring methods to mitigate reward hacking in challenging scenarios, thereby improving the model’s robustness and generalization capabilities.

C ADDITIONAL EXPERIMENTAL DETAILS

For additional experimental details, please refer to Table 4. Besides, we generate semantically analogical questions by giving rewriting prompt to Qwen3-32B. The utilized rewriting prompt is recorded as following.

```
You are given a math problem. Please rewrite it using different wording
↪ and a different real-world scenario, while keeping the underlying
↪ mathematical meaning and answer exactly the same.

Guidelines:
1. Do not change the math logic or the final answer.
2. Use different words and a new context to make it look like a different
↪ problem.
3. Avoid copying phrases or sentence structures from the original.
4. Make sure the rewritten question is natural, clear, and solvable.
5. Output ONLY between the following markers, and strictly in this format
↪ (no extra explanation):

### RESULT_START
ORIGINAL:
<original question>
REWRITE:
<rewritten question>
### RESULT_END
```

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 COMPLETE METRICS ON IFEVAL

In fact, the evaluation of IFEval includes four metrics: prompt_level_strict_acc, inst_level_strict_acc, prompt_level_loose_acc and inst_level_loose_acc. Due to space limit, we report the average of them in Table 1. We provide the complete metrics in Table 5 as a supplement. We observe that Co-Reward training not only preserves the instruction-following ability of the base models, but actually surpasses the GT reward training on Qwen2.5-3B/7B, Qwen3-4B/8B, and Llama-3.2-3B-Instruct.

Table 4: More detailed experimental parameter settings.

Training Configuration	
Train Batch Size (Number of Sampled Questions)	128
Max Prompt Length	512
Max Response Length	3072
Train Steps	170-220
Clip Ratio	0.2
Optimizer Parameters	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$)
Learning Rate	3e-06
Warmup Style	Cosine
Warmup Steps Ratio	0.1
KL Loss Coefficient	0.005
Temperature	
Training Temperature	1.0
Evaluation Temperature	0.8

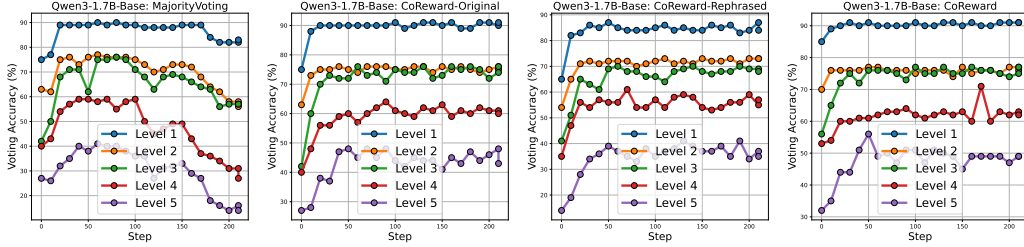


Figure 7: Voting accuracy across different difficulty level of math problems.

D.2 FURTHER VOTING ACCURACY ANALYSIS

Figure 7 illustrates the voting accuracy of Qwen3-1.7B-Base across different difficulty level as a supplement. A similar trend to Figure 5 can be observed: Majority-Voting exhibits a noticeable decline in accuracy on more challenging problems (Levels 4 and 5), whereas Co-Reward consistently maintains stable voting accuracy across all difficulty levels.

D.3 CASE STUDY FOR REPHRASED PROBLEMS

We provide several rephrased math problems in this paper, as shown below. These rephrasings clearly demonstrate that the rephrased questions offer a reasonable and high-quality alternative perspective to the original problems.

Original 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?

Rephrased Question:

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?

Original Question:

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?

Rephrased Question:

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?

Original Question:

Solve for x : $\frac{1}{2} + \frac{1}{x} = \frac{5}{6}$.

Rephrased Question:

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 .

D.4 COMPLETE GENERATION OF CASE STUDY

We provide the complete generation outputs for the case study presented in Figure 6 as a supplement. From the full outputs, we can clearly observe the reward hacking behavior exhibited by Majority-Voting, as well as the repetitive generations produced by Self-Certainty and Entropy. In contrast, both Co-Reward and GT-Reward generate the correct answer along with a well-structured reasoning path.

Original Question:

Two real numbers x and y satisfy $x - y = 4$ and $x^3 - y^3 = 28$. Compute xy .

Rephrased Question:

Two quantities a and b have a difference of 4, and the difference of their cubes is 28. Find the product ab .

Co-Reward Answer for Original Question:

Let's solve the system of equations step by step. We are given the equations:

1. $x - y = 4$ 2. $x^3 - y^3 = 28$

We start by using the identity for the difference of cubes, which is $x^3 - y^3 = (x - y)(x^2 + xy + y^2)$. Substituting the first equation $x - y = 4$ into this identity, we get:

$$x^3 - y^3 = 4(x^2 + xy + y^2)$$

We know from the problem statement that $x^3 - y^3 = 28$, so we can substitute that in:

$$4(x^2 + xy + y^2) = 28$$

Dividing both sides by 4, we get:

$$x^2 + xy + y^2 = 7$$

Next, we need to express $x^2 + y^2$ in terms of x and y . We know that:

$$x^2 + y^2 = (x - y)^2 + 2xy$$

Substituting $x - y = 4$ into the equation, we get:

$$x^2 + y^2 = 4^2 + 2xy = 16 + 2xy$$

Now substitute $x^2 + y^2 = 16 + 2xy$ into the equation $x^2 + xy + y^2 = 7$:

$$16 + 2xy + xy = 7$$

Combine like terms:

$$16 + 3xy = 7$$

Subtract 16 from both sides:

$$3xy = -9$$

Divide both sides by 3:

$$xy = -3$$

So, the value of xy is $\boxed{-3}$.

Co-Reward Answer for Augmented Question:

To solve for the product ab given the conditions $a - b = 4$ and $a^3 - b^3 = 28$, we can proceed as follows:

First, recall the identity for the difference of cubes:

$$a^3 - b^3 = (a - b)(a^2 + ab + b^2)$$

Given $a - b = 4$ and $a^3 - b^3 = 28$, substitute these into the identity:

$$28 = 4(a^2 + ab + b^2)$$

Divide both sides by 4:

$$7 = a^2 + ab + b^2$$

Next, we need to express $a^2 + ab + b^2$ in a different form. We know that:

$$a^2 + b^2 = (a - b)^2 + 2ab$$

Substitute $a - b = 4$ into the equation:

$$a^2 + b^2 = 4^2 + 2ab = 16 + 2ab$$

Substitute $a^2 + b^2 = 16 + 2ab$ back into the equation $7 = a^2 + ab + b^2$:

$$7 = 16 + 2ab + ab$$

Combine like terms:

$$7 = 16 + 3ab$$

Rearrange to solve for ab :

$$3ab = 7 - 16$$

$$3ab = -9$$

$$ab = -3$$

Thus, the product ab is $\boxed{-3}$.

GT Reward Answer:

To solve for xy given the equations $x - y = 4$ and $x^3 - y^3 = 28$, we will follow these steps:

1. **Express $x^3 - y^3$ in terms of $x - y$ and xy :

Recall the identity for the difference of cubes:

$$x^3 - y^3 = (x - y)(x^2 + xy + y^2)$$

We know from the problem that $x - y = 4$ and $x^3 - y^3 = 28$. Substituting these values into the identity, we get:

$$28 = 4(x^2 + xy + y^2)$$

Dividing both sides by 4, we obtain:

$$7 = x^2 + xy + y^2$$

2. **Express $x^2 + y^2$ in terms of $x + y$ and xy :**

We know that:

$$x^2 + y^2 = (x - y)^2 + 2xy$$

Substituting $x - y = 4$, we get:

$$x^2 + y^2 = 4^2 + 2xy = 16 + 2xy$$

Substituting this into the equation $x^2 + xy + y^2 = 7$, we get:

$$16 + 2xy + xy = 7$$

Simplifying the left side, we have:

$$16 + 3xy = 7$$

Solving for xy , we subtract 16 from both sides:

$$3xy = 7 - 16$$

$$3xy = -9$$

Dividing both sides by 3, we get:

$$xy = -3$$

Therefore, the value of xy is $\boxed{-3}$.

Majority-Voting Answer:

To solve for xy given the equations $x - y = 4$ and $x^3 - y^3 = 28$, we will follow these steps:

1. **Express $x^3 - y^3$ in terms of $x - y$ and xy :** The identity for the difference of cubes is:

$$x^3 - y^3 = (x - y)(x^2 + xy + y^2)$$

Given $x - y = 4$, we can substitute this into the identity:

$$x^3 - y^3 = 4(x^2 + xy + y^2)$$

We are also given that $x^3 - y^3 = 28$. Therefore, we can set up the equation:

$$4(x^2 + xy + y^2) = 28$$

Dividing both sides by 4, we get:

$$x^2 + xy + y^2 = 7$$

2. **Express $x^2 + y^2$ in terms of x and y :** We know that:

$$x^2 + y^2 = (x - y)^2 + 2xy$$

Substituting $x - y = 4$ into the equation, we get:

$$x^2 + y^2 = 4^2 + 2xy = 16 + 2xy$$

Now substitute $x^2 + y^2 = 16 + 2xy$ into the equation $x^2 + xy + y^2 = 7$:

$$16 + 2xy + xy = 7$$

Simplifying the left side, we get:

$$16 + 3xy = 7$$

Subtract 16 from both sides:

$$3xy = 7 - 16$$

$$3xy = -9$$

Dividing both sides by 3, we get:

$$xy = -3$$

Therefore, the value of xy is:

$$\boxed{1}$$

Self-Certainty Answer:

ex

To solve this problem, we need to follow these steps:

1. Define the variables and the equations. 2. Solve the equations to find the values of x and y . 3. Calculate the product xy .

Let's break down each step and solve the problem accordingly.

Step 1: Define the Variables and the Equations We need to define the variables x and y and the equations that represent the given conditions. The conditions are: 1. $x - y = 4$ 2. $x^3 - y^3 = 28$

Step 2: Solve the Equations We need to solve these two equations to find the values of x and y .

Step 3: Calculate the Product xy Once we have the values of x and y , we need to calculate the product xy .

Let's go through each step one by one.

Step 1: Define the Variables and the Equations 1. Define the variables x and y . 2. Write down the equations $x - y = 4$ and $x^3 - y^3 = 28$.

Step 2: Solve the Equations 1. Use a symbolic solver to solve the system of equations for x and y . 2. Extract the solutions for x and y .

Step 3: Calculate the Product xy 1. Calculate the product of the values of x and y obtained from the previous step.

Let's implement these steps in Python using sympy.

```
“python import sympy as sp
```

```
# Step 1: Define the variables x, y = sp.symbols('x y')
```

```
# Step 2: Define the equations # Equation 1: x - y = 4 eq1 = x - y - 4
```

```
# Equation 2: x3 - y3 = 28
```

```
eq2 = x**3 - y**3 - 28
```

```
# Step 3: Solve the equations to find the values of x and y # Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
```

```
# Step 4: Calculate the product xy for each solution # Extract the solutions for x and y xy_products = []
```

```
for solution in solutions:
```

```
    x_val, y_val = solution
```

```
    # Calculate the product xy
```

```
    product_xy = x_val * y_val
```

```
    # Append the product to the list of products
```

```
    xy_products.append(product_xy)
```

```
# Step 5: Output the final answer
```

```
# Since we expect only one valid solution, we can just output the first product
```

```
final_answer = xy_products[0]
```

```
# Print the final answer
```

```
print(final_answer)““
```

Step-by-Step Implementation

1. **Define the Variables and the Equations:** - Define the variables x and y . - Write down the equations $x - y = 4$ and $x^3 - y^3 = 28$.

2. **Solve the Equations:** - Use the 'solve' function from sympy to solve the system of equations for x and y .

3. **Calculate the Product xy :** - Extract the solutions for x and y . - Calculate the product xy for each solution.

4. **Output the Final Answer:** - Since we expect only one valid solution, we can just output the first product.

Now, let's implement these steps in Python.

```
“python import sympy as sp
```

```
# Step 1: Define the variables x, y = sp.symbols('x y')
```

```
# Step 2: Define the equations # Equation 1: x - y = 4
```

```
eq1 = x - y - 4
```

```
# Equation 2: x3 - y3 = 28
```

```
eq2 = x**3 - y**3 - 28
```

```
# Step 3: Solve the equations to find the values of x and y
```

```

# Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
# Step 4: Calculate the product xy for each solution
# Extract the solutions for x and y
xy_products = []
for solution in solutions:
    x_val, y_val = solution
# Calculate the product xy
product_xy = x_val * y_val
# Append the product to the list of products
xy_products.append(product_xy)
# Step 5: Output the final answer
# Since we expect only one valid solution, we can just output the first product
final_answer = xy_products[0]
# Print the final answer
print(final_answer) """
Let's go through each step in detail:
1. **Define the Variables and the Equations:** - We define the variables  $x$  and  $y$  using 'sp.symbols('x y')'. - We write down the equations  $x - y = 4$  and  $x^3 - y^3 = 28$ .
2. **Solve the Equations:** - We use the 'sp.solve' function to solve the system of equations for  $x$  and  $y$ .
3. **Calculate the Product  $xy$ : - We extract the solutions for  $x$  and  $y$  from the solutions obtained. - We calculate the product  $xy$  for each solution and store it in a list.
4. **Output the Final Answer:** - Since we expect only one valid solution, we output the first product from the list of products.
Now, let's implement this step by step in Python.
#### Step 1: Define the Variables and the Equations """python import sympy as sp
# Step 1: Define the variables
x, y = sp.symbols('x y')
# Step 2: Define the equations
# Equation 1:  $x - y = 4$ 
eq1 = x - y - 4
# Equation 2:  $x^3 - y^3 = 28$ 
eq2 = x**3 - y**3 - 28
# Print the equations for verification
print("Equation 1:", eq1)
print("Equation 2:", eq2) """
#### Step 2: Solve the Equations """python
# Step 3: Solve the equations to find the values of x and y
# Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
# Print the solutions for verification print("Solutions:", solutions) """
#### Step 3: Calculate the Product  $xy$ 
"""python
# Step 4: Calculate the product xy for each solution
# Extract the solutions for x and y
xy_products = []
for solution in solutions:
    x_val, y_val = solution
# Calculate the product xy
product_xy = x_val * y_val
# Append the product to the list of products
xy_products.append(product_xy)
# Step 5: Output the final answer # Since we expect only one valid solution, we can just output the first product
final_answer = xy_products[0]
# Print the final answer print("Final answer (product xy):", final_answer) """
Now, let's go through the implementation step by step and run the code to get the final answer.

```



```

#### Step 1: Define the Variables and the Equations “python import sympy as sp
# Step 1: Define the variables x, y = sp.symbols('x y')
# Step 2: Define the equations # Equation 1:  $x - y = 4$  eq1 = x - y - 4
# Equation 2:  $x^3 - y^3 = 28$ 
eq2 = x**3 - y**3 - 28
# Print the equations for verification
print("Equation 1:", eq1)
print("Equation 2:", eq2) “
#### Step 2: Solve the Equations
“python
# Step 3: Solve the equations to find the values of x and y
# Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
# Print the solutions for verification
print("Solutions:", solutions) “
#### Step 3: Calculate the Product  $xy$ 
“python
# Step 4: Calculate the product xy for each solution
# Extract the solutions for x and y
xy_products = []
for solution in solutions:
    x_val, y_val = solution
    # Calculate the product xy
    product_xy = x_val * y_val
    # Append the product to the list of products
    xy_products.append(product_xy)
# Step 5: Output the final answer
# Since we expect only one valid solution, we can just output the first product
final_answer = xy_products[0]
# Print the final answer
print("Final answer (product xy):", final_answer) “
Now, let’s run the code to get the final answer. “python
import sympy as sp
# Step 1: Define the variables
x, y = sp.symbols('x y')
# Step 2: Define the equations
# Equation 1:  $x - y = 4$ 
eq1 = x - y - 4
# Equation 2:  $x^3 - y^3 = 28$ 
eq2 = x**3 - y**3 - 28
# Print the equations for verification
print("Equation 1:", eq1)
print("Equation 2:", eq2)
# Step 3: Solve the equations to find the values of x and y
# Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
# Print the solutions for verification
print("Solutions:", solutions)
# Step 4: Calculate the product xy for each solution
# Extract the solutions for x and y
xy_products = []
for solution in solutions:
    x_val, y_val = solution
    # Calculate the product xy
    product_xy = x_val * y_val
    # Append the product to the list of products
    xy_products.append(product_xy)
# Step 5: Output the final answer

```

```

# Since we expect only one valid solution, we can just output the first product
final_answer = xy_products[0]
# Print the final answer
print("Final answer (product xy):", final_answer) ""
Let's break down the steps and execute the code to get the final answer.
#### Step 1: Define the Variables and the Equations 1. Define the variables  $x$  and  $y$ . 2. Write
down the equations  $x - y = 4$  and  $x^3 - y^3 = 28$ .
#### Step 2: Solve the Equations 1. Solve the system of equations to find the values of  $x$  and  $y$ .
#### Step 3: Calculate the Product  $xy$  1. Extract the solutions for  $x$  and  $y$ . 2. Calculate the
product  $xy$  for each solution.
#### Step 4: Output the Final Answer 1. Since we expect only one valid solution, output the first
product.
Let's go through each step in detail and execute the code.
##### Step 1: Define the Variables and the Equations ""python import sympy as sp
# Step 1: Define the variables x, y = sp.symbols('x y')
# Step 2: Define the equations # Equation 1:  $x - y = 4$  eq1 = x - y - 4
# Equation 2:  $x^3 - y^3 = 28$ 
eq2 = x**3 - y**3 - 28
# Print the equations for verification print("Equation 1:", eq1) print("Equation 2:", eq2) ""
##### Step 2: Solve the Equations 1. Solve the system of equations to find the values of  $x$  and  $y$ .
##### Step 3: Calculate the Product  $xy$  1. Extract the solutions for  $x$  and  $y$ . 2. Calculate the
product  $xy$  for each solution.
##### Step 4: Output the Final Answer 1. Since we expect only one valid solution, output the
first product.
Now, let's execute the code to get the final answer. ""python import sympy as sp
# Step 1: Define the variables x, y = sp.symbols('x y')
# Step 2: Define the equations # Equation 1:  $x - y = 4$ 
eq1 = x - y - 4
# Equation 2:  $x^3 - y^3 = 28$ 
eq2 = x**3 - y**3 - 28
# Print the equations for verification print("Equation 1:", eq1) print("Equation 2:", eq2)
# Step 3: Solve the equations to find the values of  $x$  and  $y$  # Solve the system of equations
solutions = sp.solve((eq1, eq2), (x, y))
# Print the solutions for verification print("Solutions:", solutions)
# Step 4: Calculate the product  $xy$  for each solution # Extract the solutions for  $x$  and  $y$  xy_products
= []
for solution in solutions: x_val, y_val = solution # Calculate the product  $xy$  product_xy = x_val
* y_val # Append the product to the list of products xy_products.append(product_xy)
# Step 5: Output the final answer # Since we expect only one valid solution, we can just output
the first product final_answer = xy_products[0]
# Print the final answer print("Final answer (product xy):", final_answer) ""
#### Step 1: Define the Variables and the Equations 1. Define the variables  $x$  and  $y$ . 2. Write
down the equations  $x - y = 4$  and  $x^3 - y^3 = 28$ .
#### Step 2: Solve the Equations 1. Solve the system of equations to find the values of  $x$  and  $y$ .
#### Step

```

Entropy Answer:

To find xy given the equations $x - y = 4$ and $x^3 - y^3 = 28$, follow these steps:

1. **Start with the first equation:**

$$x - y = 4$$

$$x - y = 4 \Rightarrow x = y + 4$$

2. **Substitute $x = y + 4$ into the second equation:**

$$x^3 - y^3 = 28$$

$$(y + 4)^3 - y^3 = 28$$

[illegible]

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$(y + 4)(y + 4) = (y + 4)(y + 4)$$

$$\setminus [(y + 4) (y + 4$$

Table 5: Other Results (%) of RL performance comparison on IFEval benchmark.

Methods	IFEval				
	Average	Prompt Strict	Prompt Loose	Inst. Strict	Inst. Loose
<i>Qwen2.5-3B</i>					
Before RL	29.83	22.55	27.17	31.89	37.70
- GT-Reward	33.66	25.51	31.42	35.85	41.85
- Self-Certainty	32.22	24.40	29.76	34.65	40.05
- Entropy	32.94	24.77	30.50	35.13	41.37
- Majority-Voting	35.19	26.25	32.72	37.53	44.24
- Co-Reward(Ours)	33.86	23.84	31.61	36.09	43.88
<i>Qwen2.5-7B</i>					
Before RL	38.19	29.57	34.57	41.85	46.76
- GT-Reward	41.49	31.79	39.56	43.65	50.96
- Self-Certainty	37.24	28.47	34.38	40.05	46.04
- Entropy	40.33	30.13	37.87	43.29	50.00
- Majority-Voting	38.60	29.21	35.86	41.61	47.72
- Co-Reward (Ours)	41.73	32.35	39.37	44.48	50.72
<i>Qwen3-1.7B-Base</i>					
Before RL	33.65	25.69	30.86	36.45	41.60
- GT-Reward	36.16	27.35	31.79	40.64	44.84
- Self-Certainty	32.96	24.58	29.20	36.69	41.36
- Entropy	35.37	26.61	31.42	39.44	44.00
- Majority-Voting	35.45	26.06	32.16	38.72	48.84
- Co-Reward (Ours)	35.56	27.91	31.23	39.32	43.76
<i>Qwen3-4B-Base</i>					
Before RL	46.43	36.04	44.18	48.68	56.83
- GT-Reward	47.80	37.34	46.77	49.40	57.67
- Self-Certainty	48.15	39.37	46.76	49.52	56.95
- Entropy	50.44	40.67	48.61	52.52	59.07
- Majority-Voting	48.78	37.89	47.50	50.36	59.65
- Co-Reward (Ours)	50.35	40.67	49.35	51.56	59.83
<i>Qwen3-8B-Base</i>					
Before RL	50.32	40.11	50.27	51.07	59.83
- GT-Reward	52.78	41.96	51.76	54.44	62.95
- Self-Certainty	50.98	39.74	49.54	52.88	61.75
- Entropy	51.81	40.67	51.20	52.76	62.59
- Majority-Voting	51.80	39.74	51.02	53.60	62.83
- Co-Reward (Ours)	55.79	43.99	57.11	55.63	66.42
<i>Llama3-2-Instruct</i>					
Before RL	57.32	46.77	55.27	60.19	67.03
- GT-Reward	47.41	37.34	42.88	52.52	57.31
- Self-Certainty	54.88	43.81	52.68	58.15	64.87
- Entropy	54.70	43.81	52.68	57.67	64.63
- Majority-Voting	47.96	37.34	43.44	52.88	58.18
- Co-Reward (Ours)	49.14	39.37	45.66	53.12	58.39