Dual Consensus: Escaping from Spurious Majority in Unsupervised RLVR via Two-Stage Vote Mechanism

Anonymous ACL submission

Abstract

Current label-free RLVR approaches for large language models (LLMs), such as TTRL and Self-reward, have demonstrated effectiveness in improving the performance of LLMs on complex reasoning tasks. However, these methods rely heavily on accurate pseudo-label estimation and converge on spurious yet popular answers, thereby trapping in a dominant mode and limiting further improvements. Building on this, we propose **D**ual **C**onsensus **R**einforcement **L**earning. (DCRL), a novel self-supervised training method which is capable of generating more reliable learning signals through a two-stage consensus mechanism. The model initially acts as an *anchor*, producing dominant responses; then it serves as an *explorer*, generating diverse auxiliary signals via a temporary unlearning process. The final training target is derived from the harmonic mean of these two signal sets. Notably, the process operates entirely without external models or supervision. Across eight benchmarks and diverse domains, DCRL consistently improves Pass@1 over majority vote while yielding more stable training dynamics. These results demonstrate that DCRL establishes a scalable path toward stronger reasoning without labels.

1 Introduction

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as an effective approach to boosting the performance of Large Language Models (LLMs), enabling superior reasoning capabilities through long chain-of-thought (Wei et al., 2023) reasoning on various challenging benchmarks (Jaech et al., 2024; DeepSeek-AI et al., 2025; Yang et al., 2025). However, typically implemented via algorithms such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024), current RLVR approaches heavily rely on human-annotated datasets or, at a minimum, environments that provide verifiable ground-truth signals (Le et al., 2022; Wang et al., 2024a). This reliance restricts its generalizability to fully unlabeled or distribution-shifted tasks where neither human annotations (Ziegler et al., 2020) nor executable environments are accessible. As LLMs approach or surpass human-level performance, they will inevitably operate in domains where even expert humans cannot provide definitive judgments or reliable evaluations, which motivates the exploration of training on unlabeled data.

Recent studies (Shafayat et al., 2025; Zhao et al., 2025a) find that LLMs can achieve self-improvement without labeled data. One approach is determinism-based methods (Prabhudesai et al., 2025; Zhang et al., 2025b), which derive rewards from the confidence of a single policy along trajectories, thereby encouraging low-entropy and high-confidence predictions. These methods can achieve better performance by sharpening the model's outputs, but it remains debatable whether they are truly effective for improving reasoning capabilities. Another approach is aggregation-based methods (Zhang et al., 2025c; Yu et al., 2025b; Wu et al., 2025), which derive rewards from agreement across multiple samples, assuming that cross-sample consistency correlates with correctness. Nevertheless, these approaches still suffer from two critical limitations (He et al.):

- **Spurious Reward Signals**: Models struggle to generate distinguishable reward signals, especially when tackling hard reasoning tasks; the answers derived from majority vote may themselves suffer from systematic biases (Zhao et al., 2025b). In the later stages of training, spurious majority outcomes can come to dominate, yet the correct solutions may instead lie within the minority rollouts.

- **Lack Exploration Capability**: By continuously rewarding consensus across diverse

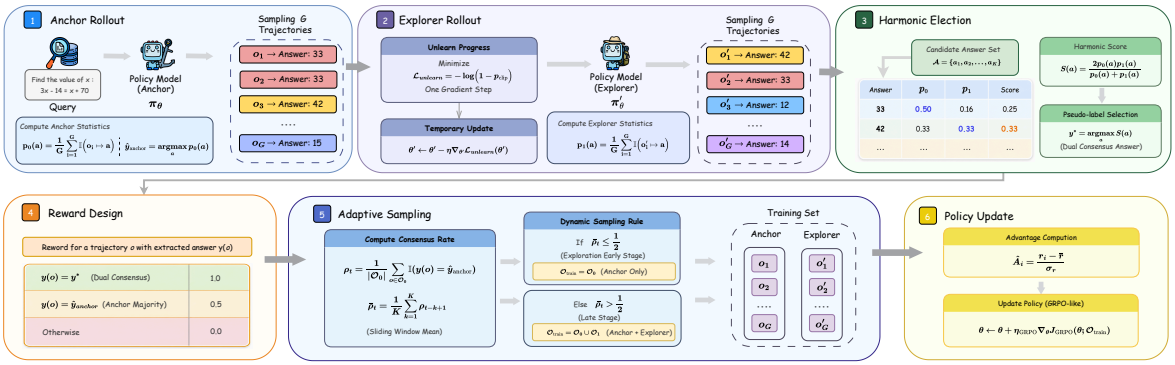


Figure 1: An overview of *Dual Consensus Reinforcement Learning* (DCRL). Specifically, the policy model assumes two roles: (1) an *anchor* that generates dominant and reliable responses; (2) an *explorer* that produces diverse auxiliary signals through a temporary unlearning process.

trajectories, the model's output distribution becomes increasingly rigid and concentrated, resulting in a severe deficiency in exploration capability or even entropy collapse (Cui et al., 2025). Consequently, the model tends to converge to a narrow set of suboptimal responses and exhibits degraded performance when confronted with out-of-domain (OOD) tasks.

In this paper, we propose *Dual Consensus*—a novel framework for Unsupervised Reinforcement Learning with Verifiable Rewards (URLVR) driven by a multi-stage vote mechanism. Our core insight stems from the following intuition: valid reasoning trajectories should not only converge to the dominant mode but also exhibit enhanced robustness when the distribution of the model is artificially flattened.

Instead of naively adopting the fragile majority vote as the pseudo-label, we decompose the rollout process into two stages: *anchor* and *explorer*. The *anchor* stage involves normal rollouts where the model generates responses under its current policy, capturing the dominant reasoning mode. Subsequently, in the *explorer* stage, we introduce a temporary unlearning process to flatten the distribution and enhance exploration, thereby encouraging the generation of diverse auxiliary responses that deviate from the dominant mode. After obtaining the two signal sets, we compute the harmonic mean of their consensus scores to determine the final reward signal. This harmonic mean balances the reliability of the dominant mode (from the anchor stage) and the diversity of potential valid trajectories (from the explorer stage), effectively mitigating the adverse impact of majority vote when it converges to spurious answers.

To validate our approach, we demonstrate the effectiveness of *Dual Consensus* through extensive experiments. We first train models on the large-scale DAPO-14K-Math dataset and evaluate them on multiple established benchmarks. Additionally, we apply test-time adaptation on five distinct datasets to further assess our method. In summary, our key contributions are as follows:

- A novel URLVR method, *Dual Consensus*, which utilizes the intrinsic robustness of the model to guide the model to evolve with policy optimization methods such as GRPO. Notably, the framework is entirely free of external models and enables continuous self-improvement without supervision.

- A pseudo-label selection mechanism and reward design that exploits the intrinsic robustness of the model itself to generate more reliable reward signals, mitigating the toxicity of majority vote when it fails.

- We empirically verify the general effectiveness of *Dual Consensus* in boosting LLMs' reasoning performance via comprehensive experiments, and additionally present systematic ablation studies and in-depth further analyses.

2 Methodology

In this section, we present the details of Dual Consensus. It mitigates spurious majority bias by generating diverse signals via Unlearn Then Explore, identifying more accurate labels through Harmonic Election, and stabilizing updates with Adaptive Sampling.

Our framework employs Grouped Relative Policy Optimization (GRPO) (Shao et al., 2024) as its

foundational RL algorithm—it stabilizes training by normalizing advantage estimates across multiple rollouts of the same prompt.

For a given input prompt x (paired with its ground-truth label y), GRPO first samples n rollouts $\{y_i\}_{i=1}^{n}$ from the current policy. For each rollout y_i, GRPO computes a reward $r_i = R(y_i, y \mid x)$, then derives the group-normalized advantage $\hat{A}_i$:

$$\hat{A}_i = \frac{r_i - \bar{r}}{\sigma_r} \tag{1}$$

where $\bar{r} = \frac{1}{n}\sum_{k=1}^{n} r_k$ denotes the mean reward of the rollout group, and σ_r is the standard deviation of the group rewards. The GRPO objective optimizes the target policy π_θ by maximizing the clipped normalized advantage:

$$\mathcal{L}_{\text{GRPO}}(x, y; \theta) = \frac{1}{n}\sum_{i=1}^{n} \min\big(\rho_i(\theta)\hat{A}_i,\ \text{clip}\big(\rho_i(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\big) - \beta\, \mathbb{D}_{\text{KL}}\big[\pi_\theta(\cdot \mid x) \parallel \pi_{\text{ref}}(\cdot \mid x)\big] \tag{2}$$

where $\rho_i(\theta) = \pi_\theta(y_i \mid x)/\pi_{\text{old}}(y_i \mid x)$ is the importance sampling ratio.

2.1 Unlearn Then Explore

Following the GRPO framework, we first initialize an ***anchor*** model (parameterized by θ') by cloning the current policy model (parameterized by θ), i.e., $\theta' \leftarrow \theta$. We then strategically apply an unlearning strategy (Liu et al., 2024) to transform this *anchor* into an ***explorer***. Unlike EEPO (Chen et al., 2025), which simply employs unlearning as a supplementary technique to enhance exploration, our approach leverages unlearning as a core methodology to actively search for correct answers.

To implement this unlearning strategy, we firstly introduce the standard *negative log-likelihood* (NLL) loss that is defined as:

$$\mathcal{L}_{\text{NLL}} = -\log \pi_{\text{anchor}}(y_{i,t} \mid x, y_{i,<t}) \tag{3}$$

This loss function imposes heavier penalties on predictions with low probability. Conversely, we take a complementary loss function that inverts the penalty pattern of the NLL loss.

To ensure numerical stability when $\pi_{\text{anchor}}(y_{i,t} \mid x, y_{i,<t}) \to 1$ (which would make $1 - \pi_{\text{anchor}}(y_{i,t} \mid x, y_{i,<t})$ approach zero and cause numerical overflow), we first perform a clipping operation on the prediction probability from the *anchor* model:

$$p_{\text{clip}} = \text{clip}\big(\pi_{\text{anchor}}(y_{i,t} \mid x, y_{i,<t}),\ \epsilon,\ 1-\epsilon\big) \tag{4}$$

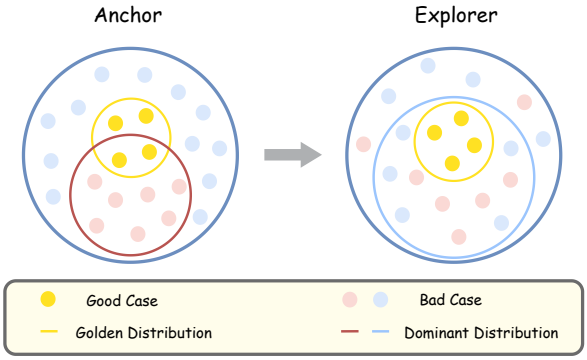


Figure 2: Output distributions of the *Anchor* and *Explorer* models. The *Explorer* model, after the unlearning process, generates a more diverse distribution.

where ϵ is a small positive constant for numerical stability. This clipping operation both avoids the term $1 - \pi_{\text{anchor}}(y_{i,t} \mid x, y_{i,<t})$ from becoming excessively small and eliminates unnecessary penalties for tokens with extremely low probabilities that are irrelevant to meaningful exploration.

Based on the clipped probability p_{clip}, we define the stabilized unlearning loss as follows:

$$\mathcal{L}_{\text{unlearn}} = -\log\big(1 - p_{\text{clip}}\big) \tag{5}$$

This loss function achieves the targeted penalty characteristic: high p_{clip} tokens lead to a large negative $\log(1-p_{\text{clip}})$, and minimizing the loss enforces a sharp reduction in their probabilities with strong penalties.

To implement this unlearning process in practice, we apply a single gradient descent step to the *anchor* model using the unlearn loss $\mathcal{L}_{\text{unlearn}}$, which transforms it into an *explorer* model:

$$\theta' \leftarrow \theta' - \eta \nabla_{\theta'} \mathcal{L}_{\text{unlearn}}(\theta') \tag{6}$$

where η denotes the learning rate for this unlearning step. Critically, this update is *temporary*—it is confined exclusively to the *anchor* model within the current iteration, and the parameters of the original policy model θ remain unchanged.

The gradient update direction $-\eta \nabla_{\theta'} \mathcal{L}_{\text{unlearn}}$ suppresses high-confidence tokens from the anchor model. The derivative of $\mathcal{L}_{\text{unlearn}} = -\log(1-p_{\text{clip}})$ assigns larger gradient magnitudes to tokens with higher p_{clip} values, and gradient descent directly drives the reduction of their probabilities to achieve the unlearning effect.

As illustrated in Fig. 2, the explorer model modulates the anchor's probability distribution, expanding the coverage of generated responses to include

trajectories that deviate from the dominant mode, which provides the necessary diverse reasoning signals for subsequent consensus aggregation.

2.2 Harmonic Election

Instead of relying on majority vote, we employ the harmonic mean to establish consensus between the anchor and explorer models, which effectively balances exploration and exploitation (shown in Fig 3). This consensus process proceeds in three sequential steps.

First, we perform anchor rollout: drawing G trajectories from the anchor policy π_{anchor}, denoted as:

$$O_0 = \{o_1, o_2, \ldots, o_G\}. \tag{7}$$

Then we apply temporary unlearning to π_{anchor} by minimizing the unlearning loss (Eq. 5), thereby suppressing its dominant generation patterns and converting it into the explorer model π_{explorer}.

We subsequently conduct explorer rollout: generating another G trajectories from π_{explorer}, forming the set:

$$O_1 = \{o'_1, o'_2, \ldots, o'_G\}. \tag{8}$$

Let $\mathcal{A}$ be the set of all candidate answers. For each $a \in \mathcal{A}$, we compute its empirical occurrence probabilities in O_0 and O_1:

$$p_0(a) = \frac{1}{G}\sum_{i=1}^{G} \mathbb{I}\big(o_i \mapsto a\big), \tag{9}$$

$$p_1(a) = \frac{1}{G}\sum_{i=1}^{G} \mathbb{I}\big(o'_i \mapsto a\big), \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

The consensus pseudo-label y^* is then selected as the answer that maximizes the harmonic mean of $p_0(a)$ and $p_1(a)$:

$$y^* = \arg\max_{a\in\mathcal{A}} \frac{2p_0(a)p_1(a)}{p_0(a)+p_1(a)}. \tag{11}$$

We note *Self-Harmony* (Wang et al., 2025) also uses the harmonic mean, yet our method only processes the model's distribution for the same input, thus avoiding Self-Harmony's semantic inconsistency in question rephrasing—an frequently recurring issue with catastrophic consequences.

During reward computation, a full reward is assigned to trajectories generating the consensus pseudo-label y^*. Trajectories consistent with the majority result of the anchor are assigned a modest reserved reward to avoid negative advantages in subsequent estimation, since such trajectories are still deemed promising relative to other counterparts. Formally, the reward for trajectory i is defined as:

$$r_i = \begin{cases} 1 & \text{if } y_i(o) = y^*, \\ 0.5 & \text{if } y_i(o) = \hat{y}_{\text{anchor}}, \\ 0 & \text{otherwise}, \end{cases} \tag{12}$$

where $\hat{y}_{\text{anchor}} = \arg\max_{a\in\mathcal{A}} \sum_{o_i\in O_0} \mathbb{I}\big(o_i \mapsto a\big)$ denotes the majority answer from the anchor trajectories.

2.3 Adaptive Sampling

Unlike the static sampling strategy in majority vote (Zuo et al., 2025), we introduce the *consensus rate* as a signal to adaptively regulate the contribution of *anchor* and *explorer* rollouts during policy updates.

We formally define the consensus rate ρ_t at step t as the proportion of anchor-sampled trajectories whose generated answers are consistent with the majority-voted result $\hat{y}_{\text{anchor}}$ of the anchor stage, which is formulated as:

$$\rho_t = \frac{1}{|O_0|}\sum_{o\in O_0} \mathbb{I}\big(y(o) = \hat{y}_{\text{anchor}}\big) \tag{13}$$

where $y(o)$ denotes the answer extracted from trajectory o, and $\mathbb{I}(\cdot)$ is the indicator function.

To capture long-term consistency and mitigate step-wise noise, we maintain a sliding window of the consensus rate over the most recent K steps and compute its mean:

$$\bar{\rho}_t = \frac{1}{K}\sum_{k=1}^{K} \rho_{t-k+1}. \tag{14}$$

A high $\bar{\rho}_t$ indicates that the policy model consistently converges to the same answer, reflecting strong model certainty and deterministic behavior; conversely, a low $\bar{\rho}_t$ suggests ongoing exploration.

Based on $\bar{\rho}_t$, we design a dynamic sampling selection rule with threshold $1/2$:

- When $\bar{\rho}_t \leq \frac{1}{2}$, only anchor trajectories O_0 are used for policy updates. Explorer trajectories O_1 still participate in consensus formation via harmonic vote but are excluded from gradient computation. This prevents premature incorporation of noisy exploratory signals while preserving rare answers that align with the pseudo-label.

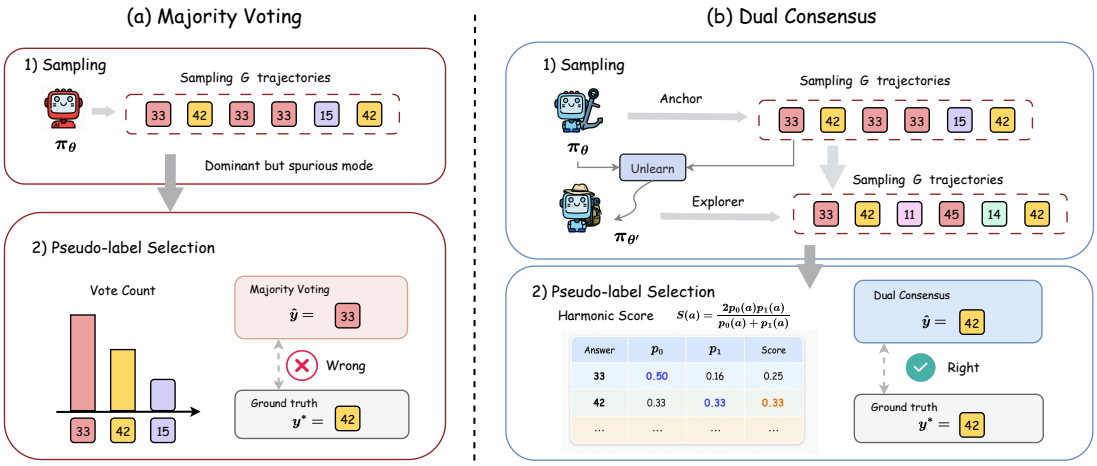


Figure 3: Comparison between *Majority vote* and *Dual Consensus*: Majority vote tends to fall into spurious consensus by over-relying on dominant but potentially incorrect response modes, while *Dual Consensus* mitigates this issue by converting the anchor model (which captures dominant reasoning patterns) into an explorer model via temporary unlearning. This transformation enables the framework to explore diverse alternative response modes, thereby balancing the reliability of current dominant patterns and the diversity of potential valid alternatives, and ultimately achieving more accurate answer selection.

- When $\bar{\rho}_t > \frac{1}{2}$, both O_0 and O_1 are included in training. This enables the policy to leverage reliable anchor behaviors while actively integrating diverse, high-quality explorations.

The effective rollout set for policy update is defined as:

$$\mathcal{O}_{\text{train}} = \begin{cases} O_0 & \text{if } \bar{\rho}_t \leq \frac{1}{2}, \\ O_0 \cup O_1 & \text{if } \bar{\rho}_t > \frac{1}{2}. \end{cases} \tag{15}$$

Only trajectories in $\mathcal{O}_{\text{train}}$ contribute to the policy gradient, ensuring a smooth transition from exploitation-dominant to balanced exploration-exploitation learning.

The overall workflow of the proposed Dual Consensus algorithm is summarized in Algorithm 1.

3 Experiments

3.1 Setups

To comprehensively evaluate the effectiveness of DCRL, our experiments are conducted under two distinct training paradigms:

- **Large-Scale Unsupervised Learning:** Directly applying DCRL to train models from scratch on a large, unlabeled dataset.
- **Test-Time Adaptation (TTA):** Using DCRL to adapt a pre-trained model to new, unseen benchmarks via constant unsupervised training.

Models: Our target models include Llama3.2-3B-Instruct (Grattafiori et al., 2024), Qwen3-4B-Base, and Qwen3-8B-Base (Yang et al., 2025). Additionally, for the TTA paradigm, we also conduct experiments on Qwen2.5-Math-1.5B (Yang et al., 2024).

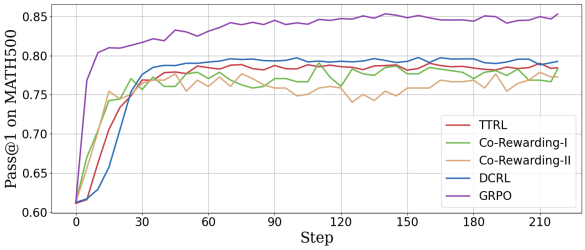
Implementation: Our primary training dataset is DAPO-Math-14k, a processed version of DAPO-Math-17k (Yu et al., 2025a). We train each model on this dataset for two epochs to mitigate overfitting. All experiments are based on the VeRL framework (Sheng et al., 2025). Implementation details are reported in Appendix A.

Benchmarks: Our benchmark suite comprises eight challenging datasets, including six math-specific: (1) MATH-500 (Hendrycks et al., 2021), (2) GSM8K (Cobbe et al., 2021), (3) AIME24 (Yang et al., 2025), (4) Minerva-math (Lewkowycz et al., 2022), (5) AMC (Zuo et al., 2025), (6) OlympiadBench (He et al., 2024); and two multi-task benchmarks: (1) MMLU-Pro (Wang et al., 2024b), (2) GPQA-Diamond (Rein et al., 2024). For the TTA experiments, we evaluate on subsets of MATH-500, AIME24, AMC, and GPQA-Diamond.

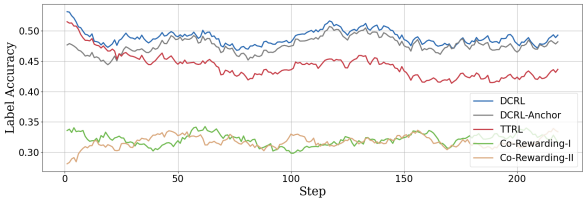
Baselines: We compare DCRL against four unsupervised RLVR methods, including one determinism-based approach **RENT** (Prabhudesai et al., 2025), and three aggregation-based approaches **TTRL** (Zuo et al., 2025), **Co-Rewarding-I**, and **Co-Rewarding-II** (Zhang et al., 2025c).

Methods	Math						Multi-Task		Average
	Math500	GSM8K	AIME24	Minerva.	AMC	Olympiad.	MMLU.	GPQA.	
Llama3.2-3B-Instruct									
Vanilla	42.4±1.17	76.1±0.79	4.5±3.82	11.7±1.98	20.6±7.08	14.6±1.09	26.4±0.06	22.3±2.97	27.3
GRPO	49.2±1.04	79.7±0.68	13.5±4.21	14.9±1.76	23.2±6.45	15.3±1.23	32.8±0.06	23.8±2.64	31.5
RENT	45.4±1.31	78.5±0.85	9.4±4.03	11.2±2.14	20.9±7.82	15.2±1.15	30.3±0.05	22.7±3.11	29.2
TTRL	44.6±1.25	71.8±0.92	10.2±2.95	12.4±1.87	21.5±5.53	14.6±1.18	25.9±0.06	21.4±2.85	27.8
Co-Rewarding-I	45.3±1.22	77.4±0.74	**11.8**±5.23	15.1±1.91	22.2±6.71	14.8±1.12	**33.9**±0.06	24.3±2.78	30.6
Co-Rewarding-II	47.6±1.19	76.9±0.81	11.6±3.89	13.7±2.05	**23.3**±4.94	15.3±1.07	24.9±0.04	21.4±3.02	29.3
DCRL(ours)	**47.8**±1.21	**79.1**±0.76	**11.8**±4.07	**15.8**±1.93	**23.3**±5.72	**15.4**±1.14	**33.9**±0.06	**24.7**±2.81	**31.4**
Qwen3-4B-Base									
Vanilla	47.4±1.07	87.0±0.57	9.3±5.09	24.5±2.01	38.3±5.49	34.9±0.83	48.6±0.15	30.7±2.72	40.0
GRPO	76.8±0.93	91.8±0.48	13.1±4.67	32.9±1.85	45.2±5.12	36.1±0.91	52.2±0.12	34.8±2.54	47.8
RENT	72.1±1.18	85.0±0.63	10.2±5.23	23.7±2.18	44.3±6.76	36.6±0.95	48.6±0.14	32.2±2.89	44.0
TTRL	74.4±0.82	91.3±0.52	11.5±3.84	29.0±1.93	44.7±6.38	37.0±0.88	51.2±0.13	32.8±2.61	46.4
Co-Rewarding-I	**74.8**±0.99	91.6±0.54	11.8±5.92	31.1±1.89	43.1±5.61	36.9±0.86	52.2±0.14	32.9±2.68	46.7
Co-Rewarding-II	74.5±1.14	**91.7**±0.41	11.6±7.01	30.8±1.96	44.5±5.42	37.1±0.89	**52.6**±0.13	33.3±2.75	47.0
DCRL(ours)	**74.8**±0.91	**91.7**±0.43	**12.5**±4.86	**31.2**±1.92	**45.6**±5.41	**37.2**±0.90	**52.6**±0.13	**34.3**±2.69	**47.4**
Qwen3-8B-Base									
Vanilla	61.8±0.96	88.3±0.46	12.5±3.94	25.0±1.62	48.9±5.66	40.1±1.47	50.6±0.15	32.3±2.92	44.9
GRPO	82.0±0.87	93.8±0.39	20.4±4.56	34.1±1.48	57.0±4.92	45.8±1.35	57.5±0.13	38.1±2.67	53.5
RENT	75.3±1.05	86.5±0.52	13.1±4.12	25.4±1.74	49.7±5.83	40.3±1.52	52.5±0.16	34.5±3.05	47.1
TTRL	78.3±0.98	92.9±0.43	14.4±3.79	**32.7**±1.85	51.2±5.41	**40.9**±1.42	55.2±0.14	36.5±3.51	50.2
Co-Rewarding-I	78.2±1.01	92.7±0.54	12.2±6.05	32.0±2.18	51.2±6.32	40.7±1.04	**56.7**±0.14	**37.9**±2.76	50.2
Co-Rewarding-II	77.2±0.83	92.9±0.45	12.2±4.07	32.3±2.31	51.8±4.58	40.2±1.49	56.1±0.15	37.6±2.84	50.0
DCRL(ours)	**79.2**±0.99	**93.3**±0.42	**14.7**±5.23	**32.7**±2.07	**51.9**±5.48	**40.9**±1.25	**56.7**±0.14	**37.9**±2.83	**50.9**

Table 1: **Main Results (%) of DCRL Trained on DAPO-Math-14k: DCRL Outperforms All Label-Free Baselines.** Results are reported as mean±std over 16 independent sampling runs. The best results are highlighted in **bold**. DCRL exceeds all unsupervised methods and nearly matches GRPO with gold labels.



(a) Accuracy curve on the MATH500 benchmark.



(b) Label accuracy curve (smoothed).

Figure 4: Training Dynamics of Dual Consensus on Qwen3-8B-Base. DCRL-Anchor in Fig. 4b refers to the majority vote of the anchor model in DCRL.

Specifically, for Co-Rewarding-I, we adopt a dataset rephrased by Qwen3-32B for all related experiments.

Metrics: We use the Pass@1 metric. For each question, we sample 16 predictions using a temperature of 0.6 and a top-p value of 0.95. The final reported score is the average Pass@1 accuracy across these 16 independent seeds.

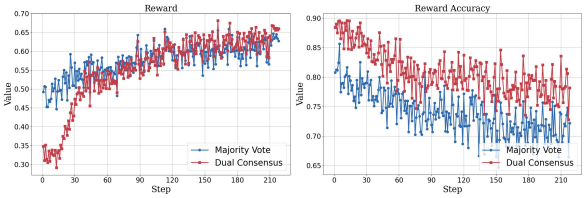


Figure 5: Comparison of reward signal between Majority Vote and Dual Consensus.

3.2 Results

3.2.1 Main Performance of Dual Consensus

Table 1 presents the main results of DCRL trained on DAPO-Math-14k across eight challenging reasoning benchmarks. Our method consistently outperforms all label-free baselines. Notably, on the Qwen3-8B-Base model, DCRL achieves 79.2% on MATH-500, surpassing the strongest baseline TTRL at 78.3% by 0.9%, and improves AIME24 from 14.4% to 14.7%, demonstrating its effectiveness on extremely hard competition-level problems. On multi-task benchmarks, DCRL matches or exceeds Co-Rewarding-I on MMLU-Pro and GPQA-Diamond, confirming its generalizability beyond pure math reasoning.

Remarkably, despite being fully unsupervised and using no ground-truth labels, DCRL achieves performance on par with, and occasionally exceeds, the supervised GRPO baseline. On Llama3.2-3B-

Algorithm 1 Dual Consensus: An Unsupervised RLVR Algorithm

Initialize: policy θ^0; learning rates $\eta_{\text{GRPO}}, \eta_{\text{u}}$; group size G; iterations T
for $t = 0$ to $T-1$ **do**
  Sample query $x \sim \mathcal{D}$
  Sample G trajectories $O_{\text{anchor}} \sim \pi_{\theta^t}(\cdot \mid x)$, compute consensus rate ρ_t and $\bar{\rho}_t$
  $\theta_{\text{e}} \leftarrow \theta^t - \eta_{\text{u}} \nabla_{\theta^t} \mathcal{L}_{\text{u}}(O_{\text{anchor}})$
  Sample G trajectories $O_{\text{explorer}} \sim \pi_{\theta_{\text{e}}}(\cdot \mid x)$
  $S(a) = \frac{2p_0(a)p_1(a)}{p_0(a)+p_1(a)}$ for each answer a in $O_{\text{anchor}} \cup O_{\text{explorer}}$
  Pseudo-label $y^* = \arg\max_a S(a)$
  if $\rho_t > 1/2$ **then**
    $O \leftarrow O_{\text{anchor}} \cup O_{\text{explorer}}$
  else
    $O \leftarrow O_{\text{anchor}}$
  end if
  Compute rewards and advantages for O
  $\theta^{t+1} \leftarrow \theta^t + \eta_{\text{GRPO}} \nabla_\theta J_{\text{GRPO}}(\theta^t; O)$
end for

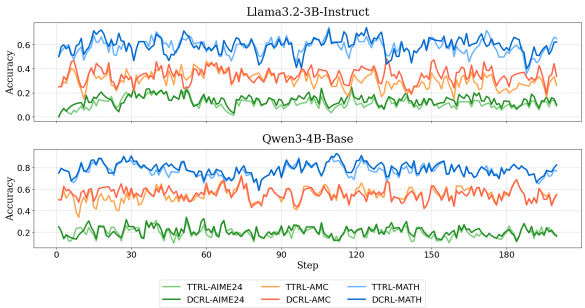


Figure 6: Label accuracy curves (smoothed) of test-time adaptation for DCRL and TTRL across different tasks.

Instruct, Qwen3-4B-Base, and Qwen3-8B-Base, it yields average gains of +7.5%, +4.0%, and +6.1% on MMLU-Pro respectively.

Compared with Majority Vote: We adopt TTRL (Zuo et al., 2025) as a representative baseline that relies on standard majority vote for pseudo-label estimation. Fig 4 and Fig 5 depict the training dynamics on Qwen3-8B-Base: our method can select low-consistency answers (especially early in training) and achieves higher ground-truth reward accuracy. This demonstrates that Dual Consensus produces more reliable supervision signals than naive majority vote.

Methods	Datasets				Average
	MATH500	AIME24	AMC	GPQA	
Qwen2.5-Math-1.5B					
Vanilla	30.6±1.28	8.5±3.70	23.4±5.46	15.4±2.30	18.7
TTRL	72.9±0.35	17.0±1.34	45.3±0.53	22.2±0.48	39.3
DCRL(ours)	74.5±0.28	17.7±1.59	46.6±0.61	22.8±0.34	40.4
Δ(-TTRL)	↑2.2%	↑4.1%	↑2.8%	↑2.7%	↑2.7%
Llama3.2-3B-Instruct					
Vanilla	42.4±1.17	4.5±3.82	20.6±7.08	22.3±2.97	22.4
TTRL	59.6±0.28	10.0±0.00	26.5±1.06	28.9±0.81	31.2
DCRL(ours)	59.8±0.19	13.3±0.00	32.3±1.52	32.6±0.64	34.5
Δ(-TTRL)	↑0.3%	↑33.0%	↑21.8%	↑12.8%	↑10.5%
Qwen3-4B-Base					
Vanilla	47.4±1.07	9.3±5.09	38.3±5.49	30.7±2.72	31.4
TTRL	82.6±0.14	17.2±1.21	56.5±2.32	35.3±0.61	47.9
DCRL(ours)	83.4±0.08	20.6±1.47	56.5±2.28	35.6±0.85	49.0
Δ(-TTRL)	↑0.9%	↑19.7%	↑0.0%	↑0.8%	↑2.2%
Qwen3-8B-Base					
Vanilla	61.8±0.96	12.5±3.94	48.9±5.66	32.3±2.92	38.8
TTRL	85.7±0.10	19.8±0.83	59.8±2.15	43.1±0.74	52.1
DCRL(ours)	86.4±0.22	22.8±1.13	61.0±1.08	44.5±0.50	53.6
Δ(-TTRL)	↑0.8%	↑15.1%	↑2.0%	↑3.2%	↑2.8%

Table 2: **Results (%) of Test-Time Adaptation Trained on Different Datasets: DCRL Consistently Outperforms the TTRL Baseline.** All results are evaluated under the same experimental settings and reported as mean±std of pass@1 over 16 independent sampling runs.

3.2.2 Test-time Adaption

Test-time Adaptation (TTA) serves as a critical validation scenario for unsupervised RLVR methods, as it directly evaluates the ability to escape spurious majority bias and generalize to unseen reasoning tasks without labeled data. The performance of TTA is presented in Table 2. We did not include methods such as *RESTRAIN* (Yu et al., 2025b) and *Self-Harmony* (Liu et al., 2025) as baselines because their official code has not been released.

Although TTA still enables generalization to other unseen scenarios with limited data (Shafayat et al., 2025; Zuo et al., 2025), a defining characteristic of TTA is its immunity to overfitting concerns. In this setting, the model's ability to a priority avoid spurious reward signals becomes critically important. As demonstrated in Table 2, our DCRL consistently outperforms TTRL across all evaluated tasks. Figure 6 shows our method effectively achieves higher label accuracy, thereby reliably validating the efficacy of our dual consensus mechanism in suppressing misleading signals.

Method	MATH	GSM8K	AIME24	MMLU.	GPQA.	Average
DCRL (Full)	**79.2**±0.99	93.3±0.42	**14.7**±5.23	56.7±0.14	**37.9**±2.83	**56.3**
- w/o Harmonic Election	78.7±1.02	**93.4**±0.43	14.3±3.98	56.2±0.14	37.8±2.98	56.0
- w/o Conservative Reward	78.3±0.88	93.3±0.49	12.2±4.21	**57.2**±0.15	36.2±3.15	55.4
- w/o Dynamic Sampling	76.4±1.15	90.3±0.50	14.3±4.16	52.0±0.15	34.4±3.14	53.4

Table 3: Ablation Studies to Analyze the Contribution of DCRL Core Modules with the Qwen3-8B-Base Model. Results are reported as mean±std over 16 independent sampling runs.

3.3 Ablation Studies

To understand the contribution of each component in our DCRL framework, we conduct a series of ablation studies on the Qwen3-8B-Base model, and the results are summarized in Table 3. More detailed results are shown in Appendix A.

Impact of Harmonic Election Replacing harmonic mean consensus with simple majority voting from the anchor model alone leads to performance drops. This confirms that harmonic election effectively mitigates spurious majority bias by fusing both dominant and diverse exploratory signals to produce more reliable pseudo-labels.

Impact of Conservative Reward Simplifying our reward design to a binary scheme (1 for correct, 0 otherwise) results in performance degradation, especially in difficult tasks. This demonstrates that our conservative reward, which reserves a modest reward for anchor majority answers, stabilizes training by preventing extreme fluctuations in advantage estimation and avoiding harsh penalties to high-confidence trajectories.

Impact of Dynamic Sampling Using both anchor and explorer samples for training at all times leads to the worst overall performance. This underscores the importance of dynamic sampling in balancing exploration and exploitation: it excludes noisy signals early on to avoid reward hacking and incorporates high-quality exploration later, ensuring stable training while preserving the ability to escape suboptimal modes.

4 Related Works

Unsupervised RL for LLMs: LLMs can achieve self-improvement without labeled data via two typical unsupervised RL paradigms: determinism-based methods (Prabhudesai et al., 2025; Zhang et al., 2025b) encourage low-entropy and high-confidence predictions for performance sharpening, while aggregation-based methods (Zhang et al., 2025c; Yu et al., 2025b; Wu et al., 2025) assign rewards by cross-sample agreement, which takes cross-sample consistency as the proxy of prediction correctness.

Test-time Adaptation for LLMs: Recent works (Akyürek et al., 2025) have demonstrated that LLMs can leverage reinforcement learning to conduct test-time adaptation (Sun et al., 2020), which effectively enhances model performance on unseen data and even surpasses the performance of standard training protocols. This paradigm (Zuo et al., 2025; Wu et al., 2025; Liu et al., 2025; Zhou et al., 2025; Zhang et al., 2025a) empowers models to dynamically adapt to novel task distributions without access to extra labeled training data.

5 Conclusion

In this paper, we propose DCRL, an unsupervised Reinforcement Learning with Verifiable Rewards (RLVR) framework that transforms intrinsic model robustness into reliable learning signals, enabling LLMs to self-improve on reasoning tasks without annotated data. By (i) adopting an unlearn strategy to break dominant suboptimal reasoning patterns and enhance exploration capability, (ii) leveraging a Harmonic Election mechanism to balance reliability and diversity for pseudo-label estimation, and (iii) introducing Adaptive Sampling to regulate the exploration-exploitation trade-off, DCRL effectively mitigates the spurious majority bias—a critical limitation of existing label-free RLVR methods. Empirically, extensive evaluations across diverse LLMs and challenging reasoning benchmarks demonstrate that DCRL consistently outperforms current determinism-based approaches and aggregation-based approaches, which paves a scalable path for LLM self-improvement without external supervision.

6 Limitations

While DCRL demonstrates significant efficacy in escaping from spurious majorities, we acknowl-

edge several critical limitations that define its operational boundaries.

Vulnerability to Joint Systematic Biases in Utterly Intractable Tasks. The foundation of DCRL relies on the premise that correct reasoning pathways possess higher intrinsic robustness than spurious ones under distribution flattening. However, when faced with out-of-distribution (OOD) or extremely complex tasks that completely transcend the capability of the base model, both the anchor policy and the explorer policy π'_θ might inherit shared, deeply ingrained systematic misconceptions. In such scenarios, the framework may suffer from a spurious consensus, where both stages consistently converge toward an identical, flawed trajectory, rendering the harmonic election mechanism ineffective.

Computational Overhead and Sampling Time. A primary constraint of DCRL is its sequential multi-stage rollout mechanism, which introduces additional computational and sampling time during each training iteration. Because the framework requires sequential generations from both the anchor and explorer policies, the data-collection phase demands more forward passes compared to traditional single-stage baselines. It is worth noting, however, that this temporal overhead does not strictly imply a multiplication of the total sample trajectory budget; the per-stage group sampling size G can be flexibly halved or scaled down to align with standard training budgets, though the sequential pipeline still incurs a non-negligible time cost.

Generalization Constraints on Free-Form Textual Domains. The mathematical design of our Harmonic Election hinges strictly upon the ability to extract deterministic, comparable final answers (e.g., numerical constants or exact options) to compute empirical consensus probabilities. Consequently, our empirical verification is heavily centered around reasoning tasks with verifiable targets. Extending DCRL to open-ended, free-form text generation fields, such as creative writing or safe alignment, remains non-trivial, as aggregating textual consensus across highly diverse semantic expressions without a precise answer pivot represents a significant methodological hurdle.

References

Ekin Akyürek, Mehul Damani, Adam Zweiger, Linlu Qiu, Han Guo, Jyothish Pari, Yoon Kim, and Jacob Andreas. 2025. The surprising effectiveness of test-time training for few-shot learning. *Preprint*, arXiv:2411.07279.

Liang Chen, Xueting Han, Qizhou Wang, Bo Han, Jing Bai, Hinrich Schutze, and Kam-Fai Wong. 2025. Eepo: Exploration-enhanced policy optimization via sample-then-forget. *Preprint*, arXiv:2510.05837.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Bingxiang He, Yuxin Zuo, Zeyuan Liu, Shangziqi Zhao, Zixuan Fu, Junlin Yang, Cheng Qian, Kaiyan Zhang, Yuchen Fan, Ganqu Cui, Xiusi Chen, Youbang Sun, Xingtai Lv, Xuekai Zhu, Li Sheng, Ran Li, Huanang Gao, Yuchen Zhang, Bowen Zhou, and 2 others. How far can unsupervised rlvr scale llm training? *Preprint*, arXiv:2603.08660.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *Preprint*, arXiv:2210.11610.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 242 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven C. H. Hoi. 2022. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Preprint*, arXiv:2207.01780.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. *Preprint*, arXiv:2206.14858.

Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. 2025. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *Preprint*, arXiv:2508.11356.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *Preprint*, arXiv:2402.08787.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *Preprint*, arXiv:2505.22660.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. 2025. Can large reasoning models self-train? *Preprint*, arXiv:2505.21444.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-time training for out-of-distribution generalization.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

Ru Wang, Wei Huang, Qi Cao, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. 2025. Self-harmony: Learning to harmonize self-supervision and self-play in test-time reinforcement learning. *Preprint*, arXiv:2511.01191.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jianghao Wu, Yasmeen George, Jin Ye, Yicheng Wu, Daniel F. Schmidt, and Jianfei Cai. 2025. Spine: Token-selective test-time reinforcement learning with entropy-band regularization. *Preprint*, arXiv:2511.17938.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Zhaoning Yu, Will Su, Leitian Tao, Haozhu Wang, Aashu Singh, Hanchao Yu, Jianyu Wang, Hongyang Gao, Weizhe Yuan, Jason Weston, Ping Yu, and Jing Xu. 2025b. Restrain: From spurious votes to signals – self-driven rl with self-penalization. *Preprint*, arXiv:2510.02172.

Haoyu Zhang, Jiaxian Guo, Yusuke Iwasawa, and Yutaka Matsuo. 2025a. Aqa-ttrl: Self-adaptation in audio question answering with test-time reinforcement learning. *Preprint*, arXiv:2510.05478.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. 2025b. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. *Preprint*, arXiv:2504.05812.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. 2025c. Co-rewarding: Stable self-supervised rl for eliciting reasoning in large language models. *Preprint*, arXiv:2508.00410.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025a. Absolute zero: Reinforced self-play reasoning with zero data. *Preprint*, arXiv:2505.03335.

Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. 2025b. The majority is not always right: Rl training for solution aggregation. *Preprint*, arXiv:2509.06870.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *Preprint*, arXiv:2509.15194.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. *Preprint*, arXiv:2506.01347.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. 2025. Ttrl: Test-time reinforcement learning. *Preprint*, arXiv:2504.16084.

A Implementation Details

A.1 Prompt

We use the same suffix prompt both in the training and evaluation of our experiments to promote clear and step-by-step reasoning:

```
\nPlease reason step by step, and put
your final answer within \boxed{}.
```

A.2 Hyperparameters

Hyperparameter settings of our experiment on Qwen3-8B-Base is shown in Table 4.

Table 4: Hyperparameter Settings for DCRL Framework.

Hyperparameter	Value
Batch size	128
Mini batch size	128
Micro batch size	8
Max prompt length	4096
Max response length	3072
Learning rate	1×10^{-6}
LR warmup steps ratio	0.1
Learning rate warmup	cosine
Optimizer	Adam
Temperature	0.6
Top k	-1
Top p	0.95
Unlearn LR	3×10^{-7}
Number of anchor samples per example	16
Number of explorer samples per example	16
Number of samples per example for policy update	16
Number of samples per example for Testing	16
Use KL loss	False

A.3 Baseline Implementation

For all baselines, we use the official code provided in their public repositories. For TTRL, we set the

learning rate to 1×10^{-6} and the warm-up ratio to 0.1 for large-scale unsupervised learning. For Co-Rewarding-I, we adopt the DAPO-Math-14k dataset rephrased by Qwen3-32B as provided in the original source code. Besides, no external models are used in all baseline experiments. All other hyperparameter settings for the baseline are kept identical to the original configuration.

B Detailed Results

B.1 Detailed Results of Ablation Studies

Detailed results of ablation studies are shown in fig 7.

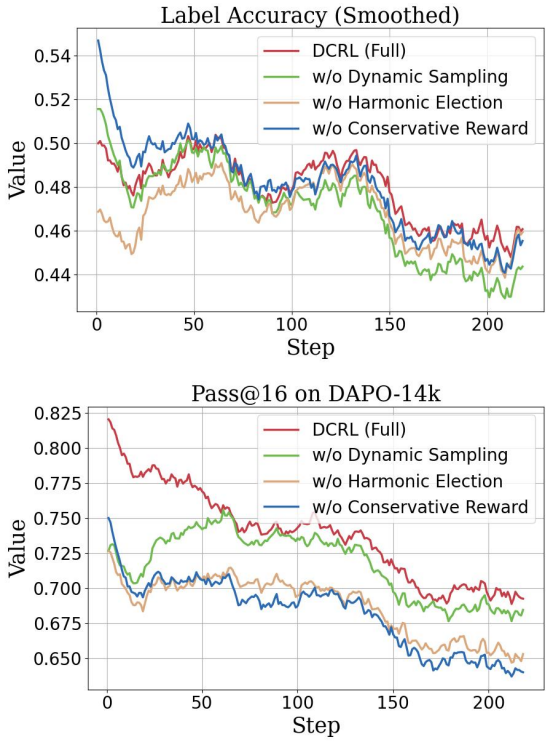


Figure 7: Detailed Results of Ablation Studies with Qwen3-8B-Base, including pass@16 on training dataset and label accuracy

C Extra Experiments

C.1 Hyperparameter Sensitivity

A key empirical finding from our experiments is that differing models necessitate tailored unlearning learning rates (η_u). A value that is too large risks disrupting the model's ability to generate valid reasoning trajectories, while one that is too small fails to effectively suppress spurious dominant modes. In this section, we present the detailed performance of the Qwen3-8B-Base model under various unlearning learning rate configurations, thereby validating the robustness of our Unlearn-Then-Explore strategy and the rationale behind our specific hyperparameter selection. Results are shown in Table 6. We also test DCRL on different models under different hyperparameters. Our experiments confirm that the method is robust to learning rate adjustments. All tested models achieve visible performance improvements and detailed results on MATH500 are shown in Table 7.

C.2 Comparison of Different Consensus Strategies

To further validate the effectiveness of our proposed Harmonic Election mechanism, we compare it against several alternative consensus strategies. These strategies differ in how they aggregate the signals from the *anchor* and *explorer* models to select the final pseudo-label y^*. The key insight is that a valid reasoning path should be robust, i.e., supported by both the dominant mode (anchor) and the exploratory distribution (explorer).

We evaluate the following strategies:

- **Majority Vote (Anchor Only)**: Simply select the majority answer from the *anchor* model's rollouts.
- **Majority Vote (Anchor + Explorer)**: A simple aggregation strategy that combines all rollouts from both the *anchor* and *explorer* models and selects the majority answer.
- **Harmonic Mean (Ours)**: Our proposed strategy, which selects the answer that maximizes the harmonic mean of its probabilities in the *anchor* and *explorer* distributions.

The results are presented in Table 5 and Fig 8. We observe that simply combining all samples (*Anchor + Explorer*) does not improve performance and can even be detrimental, as it does not effectively filter out spurious signals. In contrast, our harmonic mean strategy achieves the best overall performance, demonstrating its superior ability to balance reliability and diversity in pseudo-label selection.

D Why Does Dual Consensus Work?

We formally prove that the dual consensus pseudo-label selection mechanism achieves higher accuracy than naive majority vote by mitigating spurious majority bias, under mild and realistic assumptions.

Consensus Strategy	MATH	GSM8K	AIME24	MMLU.	GPQA.	Avg.
Majority Vote (Anchor Only)	78.7	**93.4**	14.3	56.2	37.8	56.0
Majority Vote (Anchor + Explorer)	64.7	90.3	14.3	52.0	34.4	51.5
Harmonic Mean (Ours)	**79.2**	93.3	**14.7**	**56.7**	**37.9**	**56.3**

Table 5: Comparison of Different Consensus Strategies for Pseudo-Label Selection on Qwen3-8B-Base.

Unlearn LR (η_u)	MATH	GSM8K	AIME24	MMLU.	GPQA.	Average
1×10^{-7}	78.6	92.7	12.6	55.5	37.2	55.3
3×10^{-7}	**79.2**	**93.3**	**14.7**	**56.7**	**37.9**	**56.3**
5×10^{-7}	77.6	92.7	11.5	54.3	36.4	54.5
1×10^{-6} (Failed)	67.3	88.1	12.1	50.2	35.1	50.5

Table 6: Sensitivity Analysis of Unlearning Learning Rate (η_u) on Qwen3-8B-Base.

Table 7: Performance (%) of DCRL on different models under different unlearning learning rates (Pass@1)

Model	2e-7	3e-7	5e-7
Llama3.2-3B-Instruct	47.3	47.2	47.8
Qwen3-4B-Base	72.8	74.8	74.7
Qwen3-8B-Base	78.9	79.2	78.9

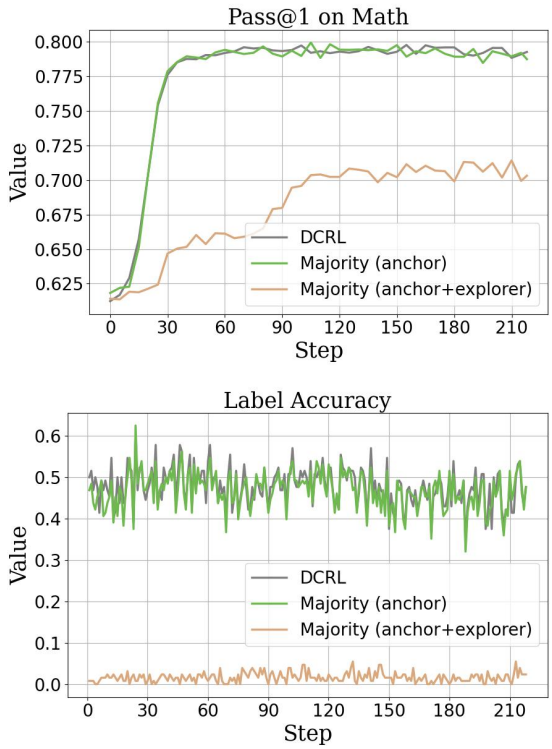


Figure 8: Curves of Different Consensus Strategies for Pseudo-Label Selection on Qwen3-8B-Base.

D.1 Problem Setup & Definitions

Let $\mathcal{A}$ be the set of candidate answers, and $y_{\text{true}} \in \mathcal{A}$ be the ground-truth answer.

- $\pi_{\text{anchor}}(a)$: Probability of answer a from the anchor model.
- $\pi_{\text{explorer}}(a)$: Probability of answer a from the explorer model (after unlearning).
- $p_0(a), p_1(a)$: Empirical probabilities of a from G rollouts of the anchor and explorer models, respectively.
- **Majority Vote**: $\hat{y}_{\text{MV}} = \arg\max_a p_0(a)$.
- **Dual Consensus**: $y^*_{\text{DC}} = \arg\max_a S(a)$, where $S(a) = \frac{2p_0(a)p_1(a)}{p_0(a)+p_1(a)}$ is the harmonic mean score.

D.2 Key Assumptions

We introduce three realistic assumptions for LLMs with spurious majority bias.

Assumption 1 (Spurious Majority Bias): There exists a spurious dominant answer $y_{\text{sp}} \neq y_{\text{true}}$ such that:

$$\pi_{\text{anchor}}(y_{\text{sp}}) \gg \pi_{\text{anchor}}(y_{\text{true}})$$

This is the core failure mode of majority vote.

Assumption 2 (Effective Unlearning): The explorer model suppresses the spurious answer but preserves the true answer:

$$\pi_{\text{explorer}}(y_{\text{sp}}) \ll \pi_{\text{anchor}}(y_{\text{sp}}),$$
$$\frac{\pi_{\text{explorer}}(y_{\text{true}})}{\pi_{\text{anchor}}(y_{\text{true}})} \gg \frac{\pi_{\text{explorer}}(y_{\text{sp}})}{\pi_{\text{anchor}}(y_{\text{sp}})}$$

The ratio inequality implies the true answer is more robust to unlearning.

Assumption 3 (Large-Sample Consistency): For sufficiently large G, by the Law of Large Numbers:

$$p_0(a) \xrightarrow{G\to\infty} \pi_{\text{anchor}}(a), \quad p_1(a) \xrightarrow{G\to\infty} \pi_{\text{explorer}}(a)$$

D.3 Main Result

Theorem: Under Assumptions 1-3, DCRL selects the true answer ($y^*_{\text{DC}} = y_{\text{true}}$), while majority vote selects the spurious answer ($\hat{y}_{\text{MV}} = y_{\text{sp}}$). Thus, $\text{Acc}(y^*_{\text{DC}}) > \text{Acc}(\hat{y}_{\text{MV}})$.

D.4 Proof

Part 1: Majority Vote Converges to Spurious Answer : By Assumption 1, $\pi_{\text{anchor}}(y_{\text{sp}}) \gg \pi_{\text{anchor}}(y_{\text{true}})$. By Assumption 3, $p_0(y_{\text{sp}}) > p_0(a), \forall a \in \mathcal{A}$. Thus $\hat{y}_{\text{MV}} = \arg\max_a p_0(a) = y_{\text{sp}}$.

Part 2: Dual Consensus Converges to True Answer : We show $S(y_{\text{true}}) > S(y_{\text{sp}})$. For large G:

$$S(a) \to \tilde{S}(a) = \frac{2\pi_{\text{anchor}}(a)\pi_{\text{explorer}}(a)}{\pi_{\text{anchor}}(a) + \pi_{\text{explorer}}(a)}.$$

Define robustness ratios (Assumption 2, $r_{\text{true}} \gg r_{\text{sp}}$):

$$r_{\text{sp}} = \frac{\pi_{\text{explorer}}(y_{\text{sp}})}{\pi_{\text{anchor}}(y_{\text{sp}})}, \quad r_{\text{true}} = \frac{\pi_{\text{explorer}}(y_{\text{true}})}{\pi_{\text{anchor}}(y_{\text{true}})},$$

where $r_{\text{true}} \gg r_{\text{sp}} \to 0$.

Substitute ratios into $\tilde{S}(a)$:

$$\begin{aligned}\tilde{S}(y_{\text{sp}}) &= \frac{2\pi_{\text{anchor}}(y_{\text{sp}}) \cdot r_{\text{sp}}\pi_{\text{anchor}}(y_{\text{sp}})}{\pi_{\text{anchor}}(y_{\text{sp}}) + r_{\text{sp}}\pi_{\text{anchor}}(y_{\text{sp}})} \\ &= \frac{2r_{\text{sp}} \cdot \pi_{\text{anchor}}(y_{\text{sp}})}{1 + r_{\text{sp}}}.\end{aligned}$$

$$\begin{aligned}\tilde{S}(y_{\text{true}}) &= \frac{2\pi_{\text{anchor}}(y_{\text{true}}) \cdot r_{\text{true}}\pi_{\text{anchor}}(y_{\text{true}})}{\pi_{\text{anchor}}(y_{\text{true}}) + r_{\text{true}}\pi_{\text{anchor}}(y_{\text{true}})} \\ &= \frac{2r_{\text{true}} \cdot \pi_{\text{anchor}}(y_{\text{true}})}{1 + r_{\text{true}}}.\end{aligned}$$

By Assumption 2, $r_{\text{sp}} \to 0 \implies \tilde{S}(y_{\text{sp}}) \to 0$. Since $\pi_{\text{anchor}}(y_{\text{true}}) > 0$ and $r_{\text{true}} > 0$, we have $\tilde{S}(y_{\text{true}}) > 0$. Thus $\tilde{S}(y_{\text{true}}) > \tilde{S}(y_{\text{sp}})$, so $y^*_{\text{DC}} = \arg\max_a S(a) = y_{\text{true}}$.

Dual Consensus enforces a robustness constraint—valid answers must be supported by both anchor and explorer, eliminating spurious answers fragile to unlearning and outperforming Majority Vote.

E The Use of Large Language Models

We used large language models (LLMs) only to polish writing and improve textual clarity. No LLM was applied to research idea generation, experimental design, data analysis, or result derivation. All scientific contributions—conceptualization, methodology, experiments, and conclusions—were independently developed by the authors in full.