Evolving Language Models without Labels: *Majority Drives Selection, Novelty Promotes Variation*

Yujun Zhou[1,2†*], **Zhenwen Liang**[1 †], **Haolin Liu**[1,3], **Wenhao Yu**[1], **Kishan Panaganti**[1], **Linfeng Song**[1], **Dian Yu**[1], **Xiangliang Zhang**[2], **Haitao Mi**[1], **Dong Yu**[1]

[1]Tencent AI Lab, [2]University of Notre Dame, [3]University of Virginia

[†] Core contributors

Correspondence to: yzhou25@nd.edu, zhenwzliang@global.tencent.com

Abstract

Large language models (LLMs) are increasingly trained with reinforcement learning from verifiable rewards (RLVR), yet real-world deployment demands models that can self-improve without labels or external judges. Existing label-free methods—confidence minimization, self-consistency, or majority-vote objectives—stabilize learning but steadily shrink exploration, causing an entropy collapse: generations become shorter, less diverse, and brittle. Unlike prior approaches such as Test-Time Reinforcement Learning (TTRL), which primarily adapt models to the immediate unlabeled dataset at hand, our goal is broader: to enable general improvements without sacrificing the model's inherent exploration capacity and generalization ability, i.e., **evolving**. We formalize this issue and propose EVolution-Oriented and Label-free Reinforcement Learning (EVOL-RL), a simple rule that couples stability with variation under a label-free setting. EVOL-RL keeps the majority-voted answer as a stable anchor (selection) while adding a novelty-aware reward that favors responses whose reasoning differs from what has already been produced (variation), measured in semantic space. Implemented with GRPO, EVOL-RL also uses asymmetric clipping to preserve strong signals and an entropy regularizer to sustain search. This "majority-for-selection + novelty-for-variation" design prevents collapse, maintains longer and more informative chains of thought, and improves both pass@1 and pass@n. EVOL-RL consistently outperforms the majority-only TTRL baseline; e.g., training on label-free AIME24 lifts Qwen3-4B-Base AIME25 pass@1 from TTRL's 4.6% to 16.4%, and pass@16 from 18.5% to 37.9%. EVOL-RL not only prevents diversity collapse but also unlocks stronger generalization across domains (e.g., GPQA). Furthermore, we demonstrate that EVOL-RL also boosts performance in the RLVR setting, highlighting its broad applicability. The code is available at https://github.com/YujunZhou/EVOL-RL.

1 Introduction

The reasoning capabilities of Large Language Models (LLMs) have advanced dramatically, particularly through paradigms like Reinforcement Learning with Verifiable Rewards (RLVR) (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025). The next frontier of autonomous intelligence lies in developing LLMs that can autonomously evolve, continuously learning from the vast, unlabeled data streams they encounter in real-world environments. This makes *label-free evolving*—letting a model improve itself while solving tasks, without ground-truth labels or external judges—both practical and necessary. However, turning inference into learning reopens a long-standing RL problem: balancing exploration and exploitation. This dilemma becomes especially severe in the label-free settings, where models must rely on internal signals, such as inherent self-consistency,

*Work done during Yujun's Internship at Tencent AI Lab.

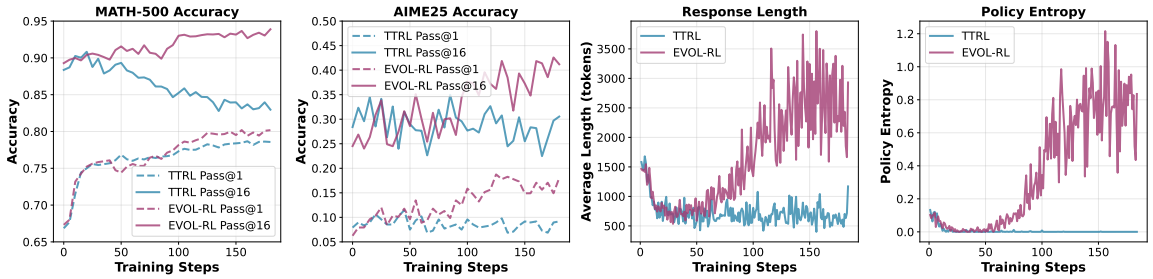

Figure 1: Illustration of entropy collapse in TTRL. The figure tracks how majority-only TTRL drives models toward collapse—pass@n deteriorates, reasoning shrinks, and entropy vanishes—while our EVOL-RL method sustains diversity and reasoning complexity. Both methods are trained in a label-free setting using the MATH-500 test set on Qwen3-4B-Base.

entropy, confidence, etc. to generate their own rewards (Grandvalet & Bengio, 2004; Lee et al., 2013; Zuo et al., 2025; Shafayat et al., 2025; Li et al., 2025b).

Although these signals can stabilize learning, but they also push the distribution to become narrower over time, leading to model degradation. This tendency reflects a critical distinction between true *evolution* and mere *test-time adaptation* (TTA). We define evolution as the ability to achieve broad-based improvements, where a model enhances its capability on the current task while maintaining or even strengthening its performance on out-of-domain (OOD) problems and preserving its overall potential (i.e., pass@k). In contrast, adaptation often leads to narrow gains on the target data at the expense of these broader capabilities. In practice, this shows up as an *entropy collapse*: uncertainty shrinks, and the model circles around a small set of familiar solutions. Figure 1 illustrates this effect on reasoning: during traditional TTRL (Zuo et al., 2025), while pass@1 may increase, pass@n drops, and the response length and complexity steadily decline, which all show that the model fails to evolve. Similar dynamics are well known in RL and self-training: without an explicit term to keep policies diverse (e.g., entropy maximization), optimization drifts toward over-confident, brittle behaviors (Haarnoja et al., 2018). Recent work also shows that self-reinforcement on self-generated data can erase the "tails of the distribution" and harm diversity over time (Shumailov et al., 2024).

In this paper, we ground LLM evolving in the simple rule behind biological evolution: *variation* creates new candidates; *selection* keeps what works. This idea motivates decades of algorithms in evolutionary computation—genetic algorithms (Holland, 1992; Eiben & Smith, 2015), novelty search (Lehman & Stanley, 2011), and quality–diversity (QD) methods such as MAP-Elites (Pugh et al., 2016)—which all show that relying on selection alone leads to premature convergence, while explicitly preserving behavioral diversity enables robust progress.

Inspired by them, we propose ***EVolution-Oriented and Label-free Reinforcement Learning (EVOL-RL)***, a simple objective that combines a stabilizing *selection* signal with an explicit *variation* incentive. Concretely, EVOL-RL retains the majority-voted answer as the anchor for stability, but adds a novelty-aware reward that scores each sampled solution by how different its reasoning is from other concurrently generated responses (semantic similarity of their reasoning traces). This "majority-for-stability + novelty-for-exploration" rule mirrors the variation–selection principle: selection prevents drift; novelty prevents collapse. As demonstrated in Figure 1, EVOL-RL successfully averts all symptoms of diversity collapse, fostering a healthy equilibrium between refining known solutions and discovering new ones. This balanced approach translates into substantial performance gains, especially in out-of-domain generalization. For instance, after training on AIME24, EVOL-RL elevates the Qwen3-4B-base model's pass@1 accuracy on the challenging AIME25 benchmark from 4.6% (TTRL) to 16.4%, while more than doubling the pass@16 accuracy from 18.5% to 37.9%.

Contributions. (1) We diagnose why majority-only objectives shrink exploration during label-free training and formalize their link to entropy collapse on reasoning tasks. (2) We provide a new

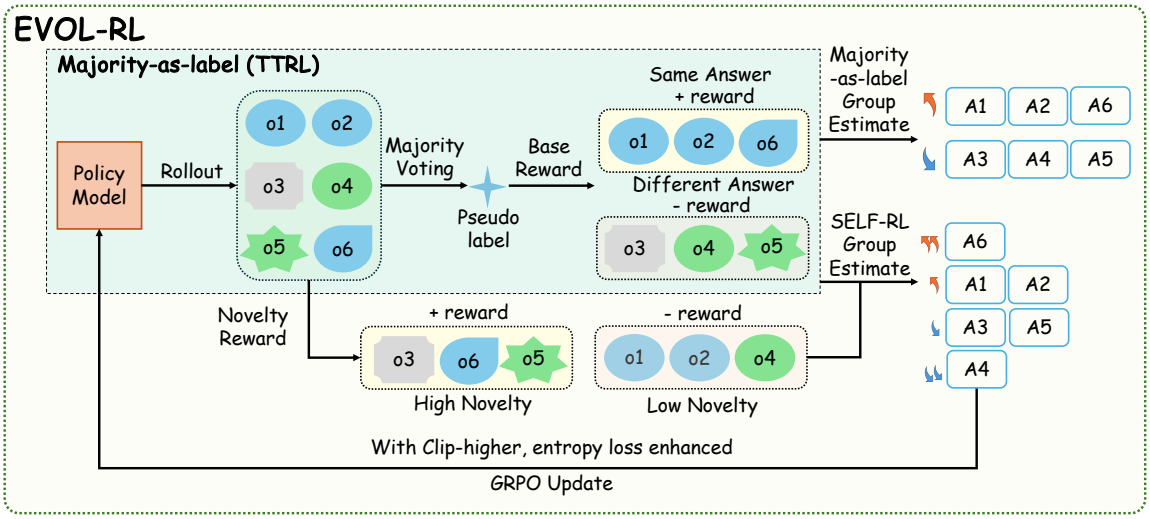

Figure 2: An overview of the EVOL-RL framework. For each prompt, the policy generates multiple responses. These are grouped by their final answer to identify the majority group. A novelty score is then computed for each response based on its semantic dissimilarity to others. Finally, a reward is assigned based on both majority (selection) and novelty (variation), guiding the policy update via GRPO. In the illustration, colors group responses by their final answer, while different marker shapes indicate semantically distinct reasoning paths.

perspective on label-free learning by framing it as an evolutionary system. This view allows us to diagnose diversity collapse as a form of premature convergence and solve it by applying the core evolutionary principle of balancing selection with variation. (3) We design a practical novelty-aware reward that complements majority selection and enables stable, label-free improvement. Across math benchmarks, EVOL-RL reverses the pass@n decline, maintains longer and more informative chains of thought, and improves out-of-domain accuracy, while remaining simple to implement. (4) We deliver state-of-the-art results in unsupervised RL training, demonstrating that by preventing collapse, EVOL-RL achieves significant out-of-domain generalization gains where prior methods fail, such as more than tripling pass@1 accuracy and doubling pass@16 accuracy on the AIME25 benchmark.

2 Method

Our goal is to create a learning framework that allows an LLM to self-improve on unlabeled data without suffering the diversity collapse inherent in self-consistency-based methods. To achieve this, we build on the Test-Time Reinforcement Learning (TTRL) paradigm but redesign its core reward mechanism to implement the principles of selection and variation. As illustrated in Figure 2, our approach, EVOL-RL, uses Generalized Reward-consistency Policy Optimization (GRPO) (Shao et al., 2024) as its optimization algorithm, but guides it with a novel reward function that explicitly balances majority with novelty.

2.1 Optimization with GRPO

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a policy-gradient algorithm designed for fine-tuning LLMs without a separate value function. Its central idea is to evaluate each sampled response relative to a group of its peers generated for the same prompt. This relative evaluation is then used to update the policy with a PPO-style clipped objective, regularized by a KL penalty to ensure stable learning.

For a given prompt $\mathbf{q}$, a policy LLM $\pi_{\theta_{\text{old}}}$ generates a group of G complete responses $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$. Each response $\mathbf{o}_i$ receives a scalar reward r_i. Rewards within the group are normalized with a z-score to obtain a response-level advantage:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)},$$

The policy is optimized with a clipped surrogate objective:

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}\hat{A}_{i,t},\ \text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right\} \tag{1}$$

Maximizing the negative of this loss increases the likelihood of responses with positive standardized advantages.

2.2 Reward Design: Implementing Selection and Variation

Our reward design directly implements the principles of selection and variation to counteract diversity collapse. Selection, based on correctness via majority vote, provides a stable signal to prevent the policy from drifting. Variation, driven by semantic novelty, provides the exploratory pressure needed to maintain a diverse set of reasoning strategies.

A key design choice is that the novelty incentive is applied strategically to all solutions—both those that agree with the majority and those that do not. For majority-aligned solutions, rewarding novelty encourages the model to discover multiple valid reasoning paths to the correct answer, directly fighting the decline in pass@n performance. For minority solutions, rewarding novelty is crucial for escaping local optima. It discourages policy collapse into a few high-frequency failure modes and instead incentivizes exploration of the broader reasoning space, which is essential for increasing the probability of discovering a previously inaccessible, correct solution path. This integration transforms the learning process: it not only mitigates diversity collapse in the current task but also aligns with the goals of continual learning. By preserving multiple reasoning modes while anchoring to a correct solution, EVOL-RL avoids forgetting potentially useful strategies and retains knowledge diversity for future tasks. Thus, training under EVOL-RL becomes not only an optimization for present performance but also a proactive investment in future adaptability.

Reward Formulation. For each prompt, the policy samples G responses $\{o_i\}_{i=1}^{G}$. Each response is scored on three criteria:

1. Validity: The response must provide a numeric final answer in a `\boxed{·}` format. Responses that fail this check are deemed invalid.

2. Majority (Selection): A binary label $y_i \in \{+1, -1\}$ is assigned based on whether a response's answer matches the majority-voted answer from the valid responses. This serves as our selection signal.

3. Novelty (Variation): We compute embeddings for the reasoning part of each response to form a cosine similarity matrix. For each response o_i, we calculate its mean similarity $\bar{s}_i$ to other responses in the same group (i.e., either majority or minority) and its maximum similarity m_i to any other response in the entire batch. The mean similarity is calculated on an intra-group basis because the majority and minority solutions are often semantically distant; a global mean would be dominated by this gap, obscuring the finer-grained variations among peer solutions within the majority group. The novelty score is:

$$u_i \;=\; 1 - \big(\alpha\,\bar{s}_i + (1-\alpha)\,m_i\big), \qquad \alpha \in \ (\text{default } 0.5).$$

This score is designed to penalize two distinct forms of redundancy: a high $\bar{s}_i$ indicates conformity to the group's semantic average, while a high m_i flags near-duplication of another specific response. The score promotes both local and global diversity. Finally, we min-max normalize the scores $\{u_i\}$

separately within the majority and minority groups to get $\tilde{u}_i$. This intra-group normalization is crucial, as it ensures that novelty is measured relative to one's direct peers, allowing for a fair comparison of diversity within each group.

Final Reward Mapping. We map the majority label and normalized novelty score into non-overlapping reward bands. This ensures that the selection signal from the majority vote is always prioritized, while novelty refines the reward within each group:

$$r_i = \begin{cases} -1, & \text{if invalid;} \\ 0.5 + 0.5\,\tilde{u}_i \in [0.5, 1], & \text{if } y_i = +1 \text{ (Majority: higher novelty earns higher reward);} \\ -1 + 0.5\,\tilde{u}_i \in [-1, -0.5], & \text{if } y_i = -1 \text{ (Minority: higher novelty mitigates penalty).} \end{cases}$$

Critically, this structure guarantees that any majority solution, regardless of its novelty, receives a higher reward than any minority solution. This maintains a strong pressure towards correctness. More details about the reward implementation are presented in Appendix A.3

Supporting Mechanisms. To further reinforce this reward design, we employ two complementary mechanisms. First, within the GRPO objective (Eq. 1), we use an asymmetric clipping range ($\epsilon_{\text{high}} > \epsilon_{\text{low}}$) (Yu et al., 2025). This allows promising and novel solutions with high advantages to receive larger gradient updates, preventing them from being prematurely clipped. Second, we add a token-level entropy regularizer to maintain diversity during the initial generation process:

$$\mathcal{L}_{\text{ent}}(\theta) = -\lambda_{\text{ent}}\, \mathbb{E}_{o\sim\pi_\theta}\left[\frac{1}{|o|}\sum_{t=1}^{|o|} H(\pi_\theta(\cdot \mid o_{<t}, x))\right], \qquad H(p) = -\sum_v p(v)\log p(v). \tag{2}$$

The total objective, $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GRPO}} + \mathcal{L}_{\text{ent}}$, thus directs learning toward semantically distinct, high-quality responses while maintaining a diverse population of solutions.

2.3 Why Majority-Only Signals Lead to Collapse and How EVOL-RL Avoids it

Why majority-only rewards cause collapse. With a binary majority-based reward, all correct (majority) responses receive the same high reward, and all incorrect (minority) responses receive the same low reward. After z-score normalization in GRPO, all majority solutions share an identical positive advantage, while all minority solutions share an identical negative one. The policy update therefore shifts probability mass uniformly toward the entire cluster of current majority solutions. Over successive updates, this process causes the probability distribution to shrink into a tight, high-confidence region. The results are precisely the symptoms observed in Figure 1: entropy drops, the model generates fewer distinct solutions, pass@n declines, and short, simplistic reasoning paths become dominant.

How EVOL-RL avoids collapse. EVOL-RL avoids this failure mode by design. The selection component (non-overlapping reward bands) ensures the model remains anchored to the high-signal majority answer. However, the variation component (the intra-group novelty score) re-orders the credit within both the majority and minority groups. Near-duplicates receive a lower reward and thus a smaller advantage, while semantically distinct solutions receive a higher reward and a larger advantage. This mechanism creates a persistent pressure against convergence to a single mode. Credit is continuously redistributed from dense clusters toward more unique solutions, preventing entropy collapse while still steering the model in the direction of correctness defined by majority.

Why a collapsed state is unstable under EVOL-RL. A collapsed state, where most generated samples are near-duplicates, is not a stable fixed point for EVOL-RL. In such a state, any sample that deviates even slightly from the dominant cluster—which will still be produced with some non-zero probability—will receive a high novelty score. This high novelty translates to a large standardized advantage, causing the GRPO update to shift probability mass away from the duplicates and toward

this more distinct sample. Therefore, a policy that produces only a single cluster of solutions cannot remain stationary; any stable policy under EVOL-RL must maintain non-zero probability across multiple, dissimilar responses.

3 Experiments

3.1 Experimental Setup

Benchmarks. To rigorously evaluate the robustness and scalability of EVOL-RL, we conduct our label-free, unsupervised training across mathematical reasoning datasets varying in size and difficulty. To test our method at scale, we use the large, standard **MATH training set (MATH-TRAIN)** (Hendrycks et al., 2021). For direct comparison with prior work and to assess performance in data-scarce, high-difficulty settings, we also follow the TTRL paradigm (Zuo et al., 2025) by training on two much smaller test sets: the general-purpose **MATH-500** and the competition-level **AIME24** (Li et al., 2024). This comprehensive setup allows us to validate EVOL-RL's versatility across both large-scale and specialized training conditions. Critically, during all training runs, we use only the problem statements, without any ground-truth labels or solutions. For evaluation, we assess the performance of our trained models on a diverse set of five benchmarks to measure both in-domain and out-of-domain generalization. The evaluation suite includes **AIME24**, **AIME25**, **MATH500**, **AMC** (Li et al., 2024), and **GPQA** (Rein et al., 2024). On these benchmarks, model-generated answers are compared against the ground-truth solutions to compute accuracy metrics.

Training Configuration. We conduct our experiments on two recent open-source base models: **Qwen3-4B-Base** and **Qwen3-8B-Base**. Our training process is implemented using the GRPO algorithm. We adopt a setup similar to that of TTRL for generating training signals. For each problem instance, we first perform a rollout phase where the policy generates 64 candidate responses. A majority label is then determined by performing a majority vote on the final answers extracted from these 64 samples. Subsequently, a random subset of 32 of these responses is used to form a batch for a single model update step. To ensure that the model has sufficient capacity for complex, multi-step reasoning, we set the maximum response length to 12,288 tokens during generation. To guide the model's reasoning process, we utilize the system prompt from SimpleRL-Zoo (Zeng et al., 2025). Implementation details are discussed in Appendix A.

3.2 Main Results

The main results of our experiments are summarized in Table 1. We highlight four key findings that demonstrate the superiority of EVOL-RL over the majority-only TTRL baseline.

EVOL-RL Enhances Both Pass@1 and Pass@16 Performance. Across all experimental settings, EVOL-RL consistently and substantially improves 'pass@16' performance over TTRL, with gains frequently exceeding 20 percentage points on the most challenging benchmarks (e.g., +24.2pp on AIME24 for the 4B model). EVOL-RL also delivers more consistent and substantial improvements to pass@1 accuracy than TTRL. Crucially, these gains are achieved while simultaneously increasing solution diversity, in sharp contrast to TTRL where pass@16 often collapses onto pass@1. This demonstrates that our method strengthens not only the model's ability to find a correct answer through multiple attempts but also its single-shot accuracy.

Consistent Improvements Across Model Scales and Training Data Sizes. The benefits of EVOL-RL are robust across both the 4B and 8B model scales, and critically, across training datasets of vastly different sizes. The performance improvements hold true whether training on the large-scale MATH-TRAIN set or the smaller, more specialized MATH-500 and AIME24 sets. This suggests that the underlying mechanism—balancing a majority anchor with novelty-driven rewards—is a fundamental improvement that scales effectively with both model capacity and data volume.

Strong Cross-Task Generalization, Bolstered by Scale. EVOL-RL demonstrates powerful generalization, learning abstract reasoning skills that transfer effectively across different mathematical

Table 1: Comparison of models trained with TTRL and EVOL-RL on the Qwen3-4B-Base and the Qwen3-8B-Base model. Each cell shows pass@1/pass@16 (averaged on 32 rollouts). Δ uses red (+) for positive and blue for negative values, showing the difference between w/EVOL-RL and w/TTRL.

Training Dataset	Model	MATH	AIME24	AIME25	AMC	GPQA
Qwen3-4B-Base						
–	Base Model	67.4/89.6	10.0/32.4	5.5/30.0	39.3/75.2	34.4/85.6
MATH-TRAIN	w/TTRL	75.4/86.9	12.1/23.2	6.8/28.6	42.5/75.2	36.5/81.4
	w/EVOL-RL	80.0/93.3	20.7/47.6	17.5/39.9	51.4/80.3	37.2/88.7
	Δ	+4.6/+6.4	+8.6/+24.4	+10.7/+11.3	+8.9/+5.1	+0.7/+7.3
MATH-500	w/TTRL	79.3/83.2	10.0/28.0	7.2/29.9	47.6/72.0	36.2/75.9
	w/EVOL-RL	79.8/93.8	19.0/43.2	16.1/41.9	50.3/82.2	38.8/89.1
	Δ	+0.5/+10.6	+9.0/+15.2	+8.9/+12.0	+2.7/+10.2	+2.6/+13.2
AIME24	w/TTRL	73.8/84.5	16.7/16.7	4.6/18.5	43.6/65.8	35.1/73.5
	w/EVOL-RL	79.6/93.6	20.6/40.9	17.1/42.0	49.9/80.9	38.0/87.8
	Δ	+5.8/+9.1	+3.9/+24.2	+12.5/+23.5	+6.3/+15.1	+2.9/+14.3
Qwen3-8B-Base						
–	Base Model	63.6/91.5	12.0/39.4	8.2/30.8	38.7/77.6	34.9/88.0
MATH-TRAIN	w/TTRL	81.1/91.1	16.7/37.6	15.6/35.9	53.6/74.0	38.1/77.1
	w/EVOL-RL	83.6/94.1	26.0/51.7	21.6/43.1	55.5/86.1	43.5/88.1
	Δ	+2.5/+3.0	+9.3/+14.1	+6.0/+7.2	+1.9/+12.1	+5.4/+11.0
MATH-500	w/TTRL	85.7/91.9	17.7/40.1	16.5/34.3	51.1/79.1	43.5/84.0
	w/EVOL-RL	84.7/95.1	24.1/49.5	20.2/44.4	58.8/86.0	43.9/92.2
	Δ	-1.0/+3.2	+6.4/+9.4	+3.7/+10.1	+7.7/+6.9	+0.4/+8.2
AIME24	w/TTRL	76.8/86.2	20.0/20.0	11.4/25.4	49.5/69.1	38.3/74.7
	w/EVOL-RL	83.1/94.2	25.4/38.1	16.5/34.7	54.4/85.8	45.2/90.0
	Δ	+6.3/+8.0	+5.4/+18.1	+5.1/+9.3	+4.9/+16.7	+6.9/+15.3

domains. A powerful example is seen with the 4B model: when trained exclusively on MATH-500, its pass@16 performance on AIME24 and AIME25 is nearly identical to the performance achieved when training on AIME24 directly, confirming that EVOL-RL learns fundamental skills rather than simply overfitting. This effect is further amplified by scale; for the 8B model, training on the large MATH-TRAIN dataset yields pass@1 performance on AIME24 (26.0%) and AIME25 (21.6%) that is far superior to training on AIME24 directly (25.4% and 16.5% respectively). This indicates that EVOL-RL effectively leverages both specialized and large-scale data to build fundamental and transferable reasoning abilities.

Robustness on Non-Mathematical Reasoning Tasks. The advantages of EVOL-RL extend beyond the domain of mathematics. On the GPQA benchmark, where TTRL consistently causes pass@16 performance to degrade compared to the base model, EVOL-RL reliably recovers and surpasses the base model. Across all training configurations, it achieves gains of +7 to +15 pp in pass@16 over TTRL, indicating that our diversity-preserving reward mechanism fosters a more generalizable reasoning ability that transfers effectively across different domains.

Table 2: Performance of Qwen3-4B-Base with EVOL-RL and its ablations on five benchmarks. Each cell reports pass@1/pass@16 accuracy.

Training Dataset	Model	MATH	AIME24	AIME25	AMC	GPQA
MATH-500	**w/EVOL-RL**	**79.8/93.8**	**19.0/43.2**	**16.1/41.9**	**50.3/82.2**	**38.8/89.1**
	-ClipHigh	75.1/91.8	12.2/31.8	11.4/31.3	42.7/73.9	32.3/81.8
	-Ent	79.5/93.4	18.3/38.5	14.7/34.3	48.3/78.6	38.6/87.0
	-ClipHigh-Ent	76.3/92.6	12.8/38.8	12.5/37.4	46.2/77.4	35.6/88.8
	-Novelty Reward	79.3/88.7	12.1/27.0	11.1/34.8	47.6/73.3	37.9/81.4
AIME24	**w/EVOL-RL**	**79.6/93.6**	**20.6/40.9**	**17.1/42.0**	**49.9/80.9**	**38.0/87.8**
	-ClipHigh	74.1/89.4	14.1/26.7	8.1/31.1	44.6/73.2	35.3/81.5
	-Ent	66.7/89.8	10.0/31.4	6.6/27.8	38.7/74.2	34.0/86.2
	-ClipHigh-Ent	75.3/89.0	16.6/26.9	9.2/32.2	45.8/71.2	37.1/82.0
	-Novelty Reward	79.4/93.0	17.7/35.6	15.9/37.4	48.8/79.6	37.9/87.1

3.3 Ablation Study

Setup. To understand the contribution of each component to EVOL-RL's performance, we conduct an ablation study on EVOL-RL-trained models on Qwen3-4B-Base. EVOL-RL introduces three key modifications compared to the TTRL baseline: (i) the novelty-aware reward function, (ii) a rollout entropy regularizer to encourage exploration, and (iii) an asymmetric PPO clipping window (higher "ClipHigh") to better preserve learning signals from high-reward samples. We systematically remove these components one at a time ("-Novelty Reward", "-Ent", "-ClipHigh") or in combination. The results are reported in Table 2.

The Critical Role of Novelty on Easier Datasets. The importance of the novelty reward is most evident when the model is trained on the MATH-500 dataset. Removing it causes the largest performance degradation in pass@16, especially on the more difficult, out-of-domain AIME24/25. This is because on a dataset with lower complexity, a majority-only approach can quickly cause the model to lock into a single, repetitive reasoning template. Our novelty reward provides the necessary counter-pressure by redistributing credit from near-duplicates to semantically different solutions, which prevents this template lock-in and promotes generalizable skills.

Exploration Mechanisms as Critical Enablers on Harder Tasks. On more challenging datasets like AIME24, where the inherent problem difficulty naturally induces a higher baseline of exploration, the other two components become more critical. In this setting, removing the entropy regularizer or the asymmetric clipping consistently lowers pass@16 performance on AIME-style problems. These mechanisms act as crucial enablers for the novelty reward: the entropy regularizer ensures a rich and continuous supply of varied reasoning paths for the novelty selector to act upon, while the higher clipping threshold preserves the full learning signal from rare but high-value solutions. They are essential for converting the naturally high exploration into stable performance gains.

Synergistic Roles for Robust Performance. Our ablations reveal a clear synergy between the three components of EVOL-RL. The novelty reward acts as the core directional selector, preventing policy collapse by re-ranking credit within the reward calculation. The entropy regularizer, in turn, sustains the exploration required during generation to provide a diverse set of candidates for the novelty mechanism to select from. Finally, the asymmetric clipping preserves the crucial learning signal from these selected high-value samples during the policy update. While the most critical bottleneck may shift depending on the dataset's complexity—from preventing template lock-in on MATH-500 to harnessing existing diversity on AIME24—it is the interplay of all three components that ensures robust and generalizable self-improvement.

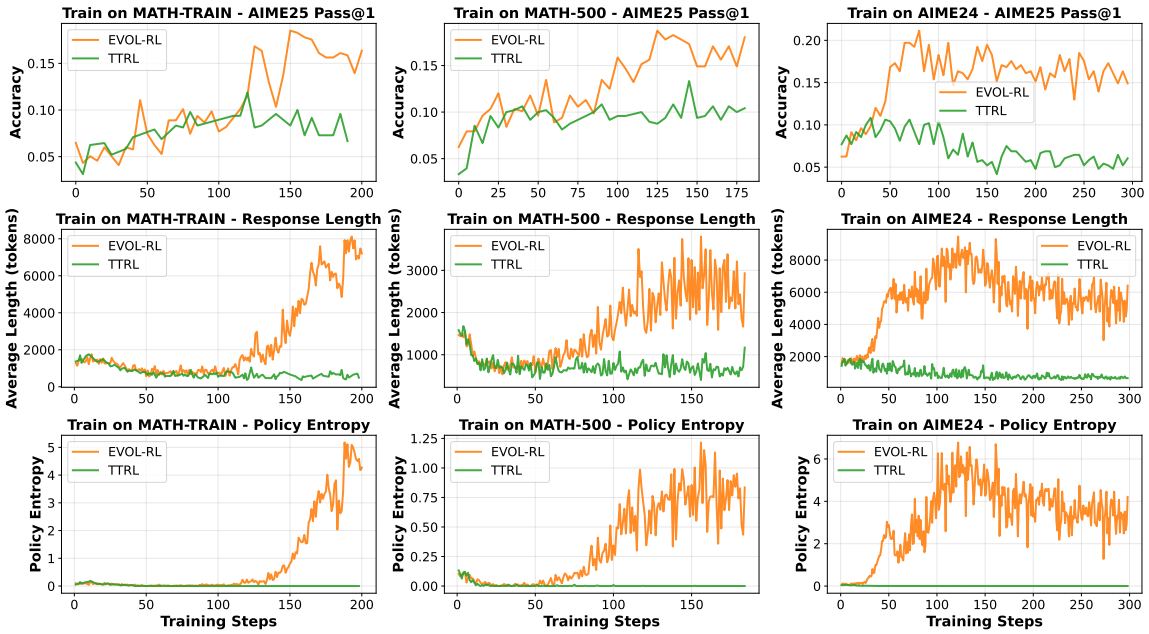

Figure 3: Training dynamics for EVOL-RL and TTRL. **Left:** models trained on *MATH-TRAIN*. **Middle:** models trained on *MATH-500*. **Right:** models trained on *AIME24*. Each panel plots, over training steps, (i) Pass@1 on *AIME25*, (ii) average response length on the training set, and (iii) policy entropy on the training set.

3.4 Training Dynamics: How EVOL-RL Escapes Entropy Collapse

To understand the reasons for EVOL-RL's better performance, we analyze its training dynamics in comparison to TTRL in a label-free setting, as shown in Figure 3. This analysis reveals how two methods, both anchored to the same majority signal, follow noticeably different paths. A consistent pattern emerges across all three training datasets, showing EVOL-RL successfully navigating away from a collapsed state that permanently traps TTRL.

Stage 1: Initial Collapse Under Majority Signal. Across all three training settings, a consistent initial dynamic unfolds: both EVOL-RL and TTRL show a sharp drop in policy entropy and average response length. This initial phase demonstrates the powerful homogenizing effect of the majority-driven reward, which quickly pushes both models toward short, high-frequency response templates. For TTRL, this collapsed state proves to be permanent; it remains trapped in this low-entropy, low-complexity state for the duration of the training run, regardless of the dataset's scale or difficulty.

Stage 2: The Evolving Point and Coordinated Recovery. Following the initial collapse, the training dynamics reveal a crucial divergence centered around a distinct **"evolving point"**. Before this point, EVOL-RL's trajectory is nearly indistinguishable from TTRL's; both models exhibit similar performance values and trends, dominated by the majority signal. However, a clear inflection point consistently emerges for EVOL-RL, after which its performance rapidly improves. While the exact timing of this "evolving point" varies across datasets, its appearance is a robust feature of our method. After this "evolving point", EVOL-RL enters a recovery phase characterized by a sustained and coordinated rise across all key metrics: policy entropy breaks away from near-zero values, average response length increases, and out-of-domain accuracy steadily climbs. This coordinated recovery allows the model to reach a new, significantly higher performance plateau where it eventually stabilizes, demonstrating its ability to break free from the majority trap.

The Mechanism of Escape. EVOL-RL's ability to escape the collapsed state comes from the synergy of its three core components. The **entropy regularizer** ensures a continuous supply of diverse rollouts, preventing the initial search space from becoming completely uniform. The **asymmetric**

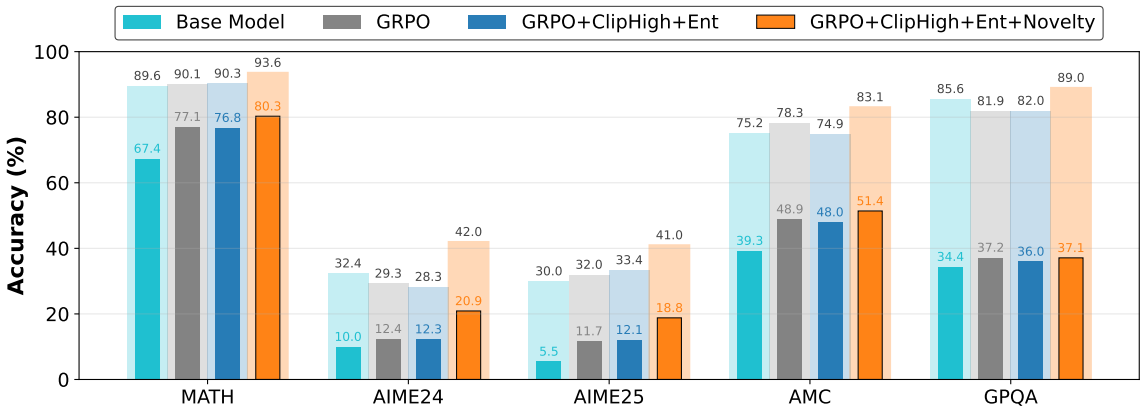

Figure 4: Performance of EVOL-RL's exploration-enhancing components when applied to a standard supervised GRPO baseline. The Qwen3-4B-Base model is trained on the MATH trainig set (Hendrycks et al., 2021) with a ground-truth verifier (RLVR).

clipping preserves the full gradient signal from the rare but high-value "majority-and-novel" samples that are crucial in the early training phase. Finally, the **novelty reward** acts as a selection pressure, consistently re-ranking credit within the majority group to favor these distinct solutions over their near-duplicate peers. Together, these mechanisms allow the model to accumulate enough exploratory signal to cross a critical threshold, shifting the policy from a single-mode, collapsed state to a multi-modal distribution that supports diverse and robust reasoning. TTRL lacks this combined mechanism and therefore remains stuck in the initial collapsed state. Furthermore, we analyze the dynamics of the majority-vote accuracy during training in Appendix B.1.

3.5 EVOL-RL Components Also Strengthen Supervised GRPO (RLVR)

Setup. We apply EVOL-RL's three exploration-enhancing ingredients to a standard supervised GRPO baseline trained on *MATH* training set (Hendrycks et al., 2021) with a ground-truth verifier (RLVR) for two epochs. Figure 4 reports the results.

Synergy Drives Generalization Gains. The primary finding is that the three components are highly synergistic, with their full combination yielding the most significant and consistent performance improvements. This complete configuration, `GRPO+ClipHigh+Ent+Novelty`, boosts pass@16 accuracy by 7% to 12% on the challenging out-of-domain AIME24 and AIME25 benchmarks. Crucially, these gains are achieved while also improving pass@1 accuracy, demonstrating that the mechanisms enhance multi-path reliability without sacrificing single-shot performance. This robust improvement extends across all evaluation benchmarks, including the cross-domain GPQA task.

Exploration and Selection are Interdependent. Analyzing the partial configurations reveals why the synergy between our components is critical. Augmenting the GRPO baseline with only the exploration mechanisms (entropy and clipping) yields inconsistent and modest gains. While these components successfully broaden the search space by encouraging varied rollouts, this exploration lacks a directional pressure to guide it toward high-quality, distinct solutions. It is only when this enhanced exploration is paired with our novelty reward—a principled selection pressure that rewards diversity—that robust and transferable reasoning gains consistently emerge across all benchmarks.

4 Related Work

Enhancing Reasoning in Large Language Models. Significant progress in LLM reasoning has been driven by RLVR (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025; Yu et al., 2025; Xiong et al.,

2025; Dai et al., 2025b), which fine-tunes models using RL on tasks where an automated verifier can confirm the correctness of the final answer, such as mathematics and coding (Zeng et al., 2025; Wang et al., 2025; Cui et al., 2025; Huang et al., 2025; Dai et al., 2025a; Zheng et al., 2025b; Zhou et al., 2025b; Zheng et al., 2025a). While highly effective, the reliance of RLVR on external verifiers restricts its applicability to domains with deterministic, easily checkable solutions (Zhao et al., 2025b; Zhou et al., 2025a; 2024). Our work contributes to the effort of improving reasoning in more general domains where such verifiers are unavailable.

Label-Free Adaptation and Self-Improvement. To overcome the limitations of verifiers and adapt to new data distributions, researchers have focused on label-free learning methods that generate reward signals without ground-truth labels. These approaches primarily fall into two categories. One line of research derives rewards from the model's **intrinsic confidence**, training the model to become more "certain" by rewarding low-entropy or self-consistent outputs (Prabhudesai et al., 2025; Agarwal et al., 2025; Zhao et al., 2025a; Zhang et al., 2025; Shafayat et al., 2025; Chung et al., 2025; Li et al., 2025a). The other prominent paradigm, which our work directly addresses, **bootstraps supervision from majority**. Test-Time Reinforcement Learning (TTRL) exemplifies this by using the majority-voted answer from multiple samples as a pseudo-label for RL updates (Zuo et al., 2025). While empirically powerful, we identify a critical flaw in the majority-driven approach: it suppresses solution diversity and actively punishes correct but non-mainstream reasoning, leading to the entropy collapse we describe. While ETTRL adjusts exploration within the original self-consistency framing (Liu et al., 2025), we are the first to pin down the majority trap and redesign the learning target to couple majority with population-level diversity, thereby removing the collapse mechanism.

5 Conclusion

In this work, we diagnose the entropy collapse, a critical failure mode in LLM evolving where majority-only rewards suppress solution diversity and harm generalization. To solve this, we propose EVOL-RL, a framework that balances the stability of majority-vote selection with an explicit variation incentive that rewards semantic novelty. Our experiments demonstrate that EVOL-RL successfully prevents collapse by maintaining policy entropy and reasoning complexity, which translates into substantial performance gains on both in-domain and out-of-domain benchmarks. By anchoring learning to a stable majority signal while simultaneously encouraging exploration, EVOL-RL offers a robust and practical methodology for enabling LLMs to continuously and autonomously evolve without external labels.

References

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.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying large language model post-training for diverse creative writing. *arXiv preprint arXiv:2503.17126*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, et al. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*, 2025a.

Runpeng Dai, Tong Zheng, Run Yang, Kaixian Yu, and Hongtu Zhu. R1-re: Cross-domain relation extraction with rlvr. *arXiv preprint arXiv:2507.04642*, 2025b.

Agoston E Eiben and James E Smith. *Introduction to evolutionary computing*. Springer, 2015.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

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.

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

John H Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Joel Lehman and Kenneth O Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary computation*, 19(2):189–223, 2011.

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, 2024.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly reinforcing diversity and quality in language model generations. *arXiv preprint arXiv:2509.02534*, 2025a.

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, et al. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*, 2025b.

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

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

Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, 3:40, 2016.

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

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

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.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, et al. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*, 2025.

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*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

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*, 2025.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025a.

Yulai Zhao, Haolin Liu, Dian Yu, SY Kung, Haitao Mi, and Dong Yu. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*, 2025b.

Tong Zheng, Lichang Chen, Simeng Han, R Thomas McCoy, and Heng Huang. Learning to reason via mixture-of-thought for logical reasoning. *arXiv preprint arXiv:2505.15817*, 2025a.

Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Xinyu Yang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, et al. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*, 2025b.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing general reasoning without verifiers. *arXiv preprint arXiv:2505.21493*, 2025a.

Yujun Zhou, Yufei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. Defending jailbreak prompts via in-context adversarial game. *arXiv preprint arXiv:2402.13148*, 2024.

Yujun Zhou, Jiayi Ye, Zipeng Ling, Yufei Han, Yue Huang, Haomin Zhuang, Zhenwen Liang, Kehan Guo, Taicheng Guo, Xiangqi Wang, et al. Dissecting logical reasoning in llms: A fine-grained evaluation and supervision study. *arXiv preprint arXiv:2506.04810*, 2025b.

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.

A Implementation Details

This section provides additional details on the implementation of our reward formulation and supporting mechanisms.

A.1 System Prompt

For all experiments, we used the following system prompt to guide the model's generation format, ensuring that it produces a step-by-step reasoning process and a clearly marked final answer (Zeng et al., 2025):

System Prompt

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

A.2 Answer and Reasoning Extraction

To implement the scoring criteria described in the main text, we apply the following extraction procedure for each generated response o_i:

- **Final Answer Extraction (for Validity):** We parse the response to find the content within the final occurrence of the `\boxed{·}` command. A response is deemed "valid" only if this command is present and its content contains at least one numeric digit. This extracted numeric string is used for the majority vote.

A.3 Novelty Score Calculation Details

The novelty score u_i relies on computing semantic similarity between the reasoning parts of the generated responses.

Embedding Model. We use the **Qwen3-4B-Embedding** model to generate dense vector representations for the extracted reasoning parts. Each vector is L2-normalized before similarity computation.

Cosine Similarity Matrix. For a group of G responses with corresponding L2-normalized embedding vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_G\}$, the cosine similarity matrix $\mathbf{S} \in \mathbb{R}^{G \times G}$ is computed as $\mathbf{S} = \mathbf{V}\mathbf{V}^T$, where $\mathbf{V}$ is the matrix whose rows are the vectors $\mathbf{v}_i$. The element S_{ij} represents the cosine similarity between the reasoning of response o_i and o_j.

Intra-Group Min-Max Normalization. To obtain the normalized novelty score $\tilde{u}_i \in [0, 1]$ from the raw scores $\{u_k\}$ within a specific group (e.g., the majority group), we apply standard min-max normalization:

$$\tilde{u}_i = \frac{u_i - \min(\{u_k\})}{\max(\{u_k\}) - \min(\{u_k\}) + \epsilon_{\text{norm}}}$$

where ϵ_{norm} is a small constant (e.g., 10^{-8}) to prevent division by zero in cases where all novelty scores in the group are identical.

A.4 Hyperparameter Settings

For our label-free experiments, we largely follow the settings established by TTRL to ensure a fair comparison. The general hyperparameters are detailed in Table 3, and the settings specific to our EVOL-RL method are listed in Table 4.

Table 3: General hyperparameters for label-free training, following TTRL.

Hyperparameter	Value
Train Batch Size	8
PPO Mini-Batch Size	1 (effective size of 32)
PPO Micro-Batch Size	2
Rollouts for Majority Vote	64
Rollouts Used for Training	32
Generation Temperature	1.0
Validation Temperature	0.6
Learning Rate	5e-7
Use KL Loss	True
KL Loss Coefficient	0.001

Table 4: Key hyperparameters specific to the EVOL-RL framework.

Hyperparameter	Value
Asymmetric Clipping High (ϵ_{high})	0.28
Entropy Regularizer Coefficient (λ_{ent})	0.003
Novelty Score Mixing Coefficient (α)	0.5

B Additional Experimental Results

B.1 Analysis of the Majority Vote Signal

To further investigate the differences between EVOL-RL and TTRL, we analyze the quality of the training signal itself by tracking the accuracy of the majority vote (maj@16) over the course of training, as shown in Figure 5. This analysis reveals how the self-generated pseudo-labels evolve under each method.

A highly consistent pattern emerges across all three training datasets. TTRL initially improves the maj@16 accuracy over the base model, but it quickly converges to a performance plateau. For the remainder of the training, its maj@16 accuracy remains largely unchanged, indicating that the consensus-only approach rapidly finds a local optimum for the consensus answer and becomes locked in, unable to discover better solutions.

In contrast, EVOL-RL exhibits a markedly different dynamic. While its initial trajectory often mirrors that of TTRL, reflecting the early stabilizing influence of the consensus signal, a clear divergence occurs. Consistent with the inflection point observed in our main training dynamics analysis, EVOL-RL's maj@16 accuracy breaks away from the TTRL plateau and begins a second, sustained ascent. It reliably climbs to and stabilizes at a significantly higher level of accuracy. This demonstrates that EVOL-RL's exploration mechanisms not only improve the final policy but also progressively refine the quality of the pseudo-labels used for training, allowing the model to escape suboptimal consensus and continuously improve its understanding of the task.

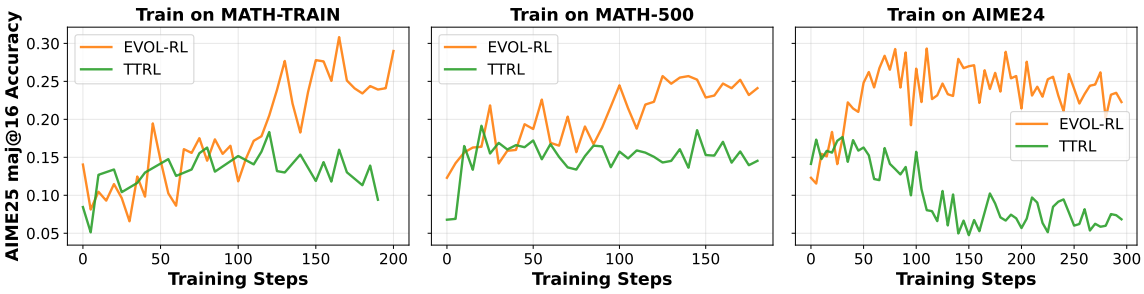

Figure 5: Training dynamics of the majority-vote accuracy (maj@16) for EVOL-RL and TTRL. Each panel plots the accuracy of the consensus answer derived from 16 rollouts over the course of training. The training datasets are: (**Left**) MATH-TRAIN, (**Middle**) MATH-500, and (**Right**) AIME24.