ON THE EVOLUTION OF LANGUAGE MODELS WITHOUT LABELS: MAJORITY DRIVES SELECTION, NOVELTY PROMOTES VARIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can improve through reinforcement learning with verifiable rewards (RLVR), but many settings lack reference answers or external verifiers. Label-free self-improvement instead constructs training signals from the model's own outputs. Because the same policy generates both the candidate responses and the signals derived from them, each update also changes the candidates available for future training, creating a self-reinforcing feedback loop. We identify a failure mode of this loop, which we call *selection-induced diversity collapse*: repeated self-selection can progressively narrow the candidate pool and make alternative correct solutions harder to sample. To address this problem, we propose EVOL-RL, which balances *selection* and *variation*. Majority voting provides a stable selection signal, while embedding-based novelty refines credit among responses with the same selection status; entropy regularization and asymmetric clipping further support this process by maintaining diverse rollouts and preserving rare, high-advantage updates. Across multiple training distributions, model scales, backbones, and matched label-free baselines, EVOL-RL consistently improves pass@n while generally maintaining or improving pass@1. For example, when trained without labels on AIME24, Qwen3-4B-Base improves over TTRL on held-out AIME25 from 4.6% to 17.1% in pass@1 and from 18.5% to 42.0% in pass@16. Training analyses further show that when the current majority is wrong, EVOL-RL better preserves alternative correct solutions. Code is available at https://anonymous.4open.science/r/EVOL-RL.

1 INTRODUCTION

Reinforcement learning with verifiable rewards has advanced the reasoning capabilities of large language models (LLMs) (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025), but continued improvement cannot always rely on reference answers or external judges. Label-free self-improvement instead derives supervision from the model's own outputs, using signals such as pseudo-labels, confidence, or consistency (Grandvalet & Bengio, 2004; Lee et al., 2013; Zuo et al., 2025; Shafayat et al., 2025; Li et al., 2025b). Concretely, the current policy generates candidate responses, these responses are used to construct rewards or pseudo-labels, and the resulting signal updates the policy. The updated policy then generates the candidates used to construct supervision in the next round. As a result, each update changes not only what the model generates next, but also the candidate solutions from which its next training signal will be constructed.

This closed loop becomes problematic when the self-generated signal repeatedly favors responses that already agree with the model's dominant belief. Those responses become more likely under the updated policy, while alternative reasoning paths are sampled less often; future supervision is then constructed from an increasingly concentrated candidate set, further reinforcing the same modes. Related concerns about self-reinforcing feedback and learning from model-generated data have been studied in self-feedback, alignment, and recursive training (Liang et al., 2024; Lee et al., 2024; Ding et al., 2024; Liang et al., 2025b; Shumailov et al., 2024; Li et al., 2026). We refer to this failure mode as *selection-induced diversity collapse*. Its central consequence is the *loss of corrective alternatives*: alternative correct solutions can become progressively less likely to be sampled, making an erroneous consensus harder to overturn even as single-sample accuracy improves.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

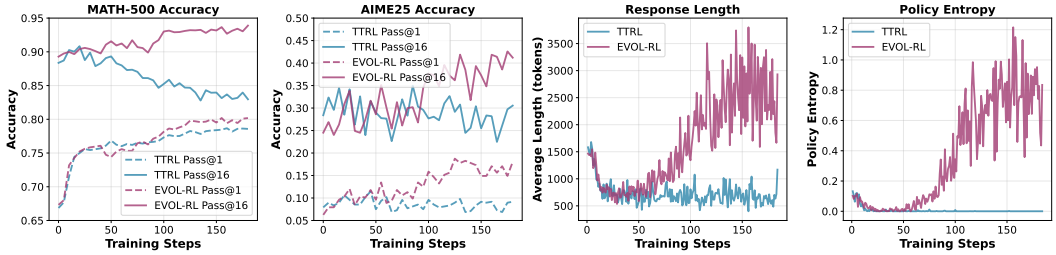

Figure 1: Training dynamics of TTRL and EVOL-RL on Qwen3-4B-Base. TTRL improves pass@1, but pass@16, response length, and policy entropy decline; EVOL-RL preserves broader coverage.

To make this failure mode concrete, we consider TTRL (Zuo et al., 2025), a representative majority-vote approach to label-free self-improvement. TTRL uses the majority answer among the current policy's rollouts as a pseudo-label and reinforces responses that agree with it. It therefore directly instantiates the feedback loop described above: the policy determines the candidate set, the candidate set determines the supervision, and that supervision is then used to update the policy. As shown in Figure 1, TTRL can improve single-sample accuracy while multi-sample coverage, response length, and policy entropy decline.

TTRL provides the clearest example, but the concern is broader: repeatedly reinforcing a model's belief can narrow the candidate set from which future supervision is constructed. In our matched comparisons, several other label-free objectives show similar pass@1–pass@k trade-offs, suggesting that gains in pass@1 do not necessarily preserve alternative solutions for future correction.

The problem is not selection itself, but selection without a mechanism to preserve diversity. Label-free self-improvement needs a stable signal that favors promising solutions, while still keeping other plausible solutions available as the policy evolves. We describe these two roles as *selection* and *variation*. This balance mirrors the classical selection–variation principle in evolutionary computation (Holland, 1992; Eiben & Smith, 2015; Lehman & Stanley, 2011; Pugh et al., 2016). In our setting, variation is valuable not only for exploration, but also for preserving the candidate pool from which future supervision is constructed.

We instantiate this principle in *EVolution-Oriented and Label-free Reinforcement Learning* (EVOL-RL). Its core design separates *selection* from *variation*: majority voting first determines which responses are favored, and an embedding-based novelty signal then redistributes credit among responses with the same majority status. This preserves variation without allowing novelty to override the selection signal. Entropy regularization and asymmetric clipping support this process by maintaining a supply of alternative rollouts and preserving updates from rare, high-advantage samples (Haarnoja et al., 2018; Yu et al., 2025). Together, these mechanisms keep alternative reasoning paths available while the model continues to learn from its own outputs.

Contributions. **(1)** We diagnose a failure mode of label-free self-improvement in which self-generated selection progressively narrows exploration, making alternative correct solutions harder to recover and eventually leading to diversity collapse. **(2)** We frame label-free self-improvement as a balance between selection and variation, instantiate this principle in EVOL-RL, and provide an idealized analysis showing that variation promotes broader coverage of correct reasoning modes. **(3)** Across multiple training settings, model scales, backbones, and matched baselines, EVOL-RL reverses the pass@n degradation observed in prior label-free methods, preserves alternative reasoning paths, and improves held-out reasoning performance.

2 Related Works

Enhancing Reasoning in LLMs. 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; Wang et al., 2025c; Zhuang et al., 2025), 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., 2025a;b; Cui et al., 2025; Huang et al., 2025; Dai et al., 2025a; Zheng et al., 2025b; Zhou et al., 2025b; Zheng et al., 2025a; Fang et al., 2025; Liu et al., 2025b). While highly effective, the reliance of RLVR on external verifiers restricts its applicability to domains with

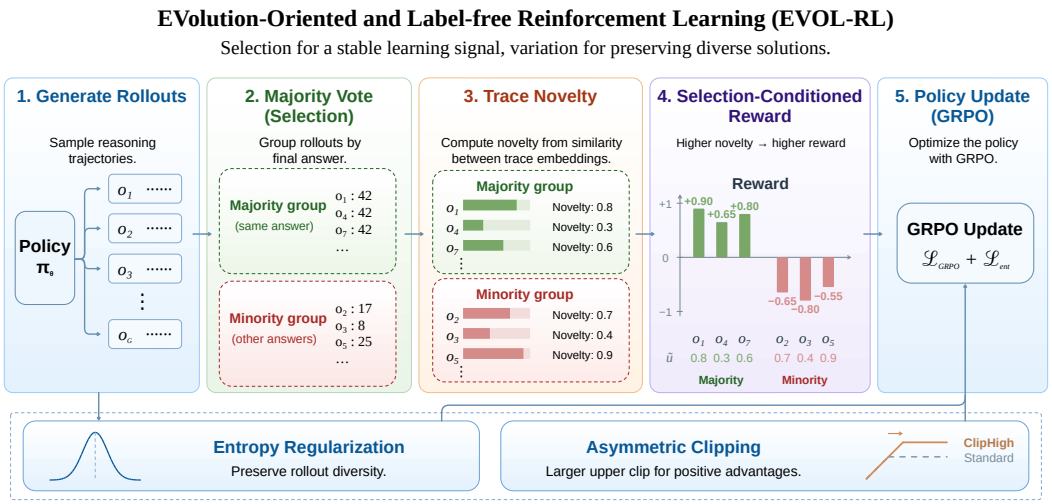

Figure 2: Overview of EVOL-RL. Majority voting provides the selection signal, while trace novelty refines rewards within each selection group. Entropy regularization maintains rollout diversity, and asymmetric clipping preserves rare, high-advantage updates. Values are illustrative.

deterministic, easily checkable solutions (Zhao et al., 2025c;a; Zhou et al., 2025a; 2024; Liang et al., 2025a). Our work instead studies self-improvement when such external verification is unavailable.

Label-Free Adaptation and Self-Improvement. To reduce reliance on external verifiers, label-free methods derive training signals from the model's own outputs using confidence, entropy, self-consistency, or agreement across multiple samples (Prabhudesai et al., 2025; Agarwal et al., 2025; Zhao et al., 2025b; Zhang et al., 2025b; Shafayat et al., 2025; Chung et al., 2025; Zhang et al., 2025a; Zuo et al., 2025). TTRL, for example, uses the majority-voted answer as a pseudo-label for RL updates (Zuo et al., 2025). Because these signals are generated by the current policy itself, repeated reinforcement can narrow the candidate pool and make alternative correct solutions harder to sample. Related work improves exploration through entropy regularization, asymmetric clipping, or other exploration objectives (Cui et al., 2025; Park et al., 2025; Liu et al., 2025a), while some approaches introduce separately trained evaluators (Li et al., 2025a; Pang et al., 2024). EVOL-RL instead explicitly separates selection from variation: majority voting determines selection status, while embedding-based novelty refines credit only among responses with the same status.

3 METHOD

Our approach is illustrated in Figure 2. EVOL-RL uses Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for policy optimization, while separating two roles in the training signal: majority voting provides selection, and trace novelty preserves variation within each selection group. Entropy regularization and asymmetric clipping support this process by maintaining diverse rollouts and preserving high-advantage updates.

3.1 OPTIMIZATION WITH GRPO

GRPO 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. For a prompt $\mathbf{q}$, a policy LLM $\pi_{\theta_{\text{old}}}$ generates a group of G 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:

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

3.2 Reward Design: Selection-Conditioned Variation

Our reward design assigns distinct roles to *selection* and *variation*. Majority voting first determines whether each response agrees with the current consensus. Novelty then refines credit *within* the majority and minority groups, rather than competing with the selection signal. Among majority responses, more distinct reasoning traces receive higher rewards than near-duplicates. Among minority responses, higher novelty leads to a less negative reward, preserving alternative trajectories without rewarding disagreement with the majority itself. Non-overlapping reward bands ensure that selection always takes priority over variation.

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

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

2. Majority (Selection). Among valid responses, we identify the majority-voted final answer and assign a binary label $y_i \in \{+1, -1\}$ according to whether response o_i agrees with it. This label determines the response's selection group.

3. Novelty (Variation). We embed the reasoning trace of each response and compute pairwise cosine similarities. For response o_i, let $\bar{s}_i$ denote its mean similarity to other responses in the same selection group, and let m_i denote its maximum similarity to any other response in the rollout set. We define

$$u_i = 1 - (\alpha \bar{s}_i + (1-\alpha) m_i), \qquad \alpha \in [0,1],$$

with $\alpha = 0.5$ by default.

The two terms capture complementary forms of redundancy: $\bar{s}_i$ measures how closely a response follows the common pattern within its group, while m_i detects near-duplicates. We then min–max normalize $\{u_i\}$ separately within the majority and minority groups to obtain $\tilde{u}_i \in [0,1]$. This keeps novelty relative to responses with the same selection status.

Final Reward Mapping. We map selection status and normalized novelty into non-overlapping reward bands:

$$r_i = \begin{cases} -1, & \text{if invalid,} \\ 0.5 + 0.5\,\tilde{u}_i \in [0.5, 1], & \text{if } y_i = +1, \\ -1 + 0.5\,\tilde{u}_i \in [-1, -0.5], & \text{if } y_i = -1. \end{cases}$$

Thus, any majority response receives a higher reward than any minority response, while novelty determines the relative credit within each group. This preserves the majority-based selection signal while discouraging concentration on near-duplicate reasoning traces. Additional implementation details are provided in Appendix A.4.

3.3 Supporting Mechanisms

The reward design determines how selection and variation allocate credit. We use two supporting mechanisms to help preserve the resulting learning signals during optimization.

First, we use an asymmetric clipping range ($\epsilon_{\text{high}} > \epsilon_{\text{low}}$) (Yu et al., 2025). A larger upper clipping threshold allows positive-advantage responses to increase in probability over a wider range before being clipped, preserving stronger updates from rare high-advantage samples.

Second, we add a token-level entropy regularizer to discourage premature policy concentration and maintain a diverse supply of rollouts for subsequent training steps. In loss notation, $\mathcal{L}_{\text{GRPO}}$ denotes the negative clipped surrogate, and we minimize the entropy loss

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

Together, asymmetric clipping preserves strong positive updates, while entropy regularization helps maintain the candidate diversity on which the selection–variation reward operates.

3.4 Theoretical Analysis

To isolate the role of variation from pseudo-label noise and optimization details, we analyze a simplified objective in which correctness is known and variation is represented by pairwise trajectory dissimilarity:

$$J_0(\pi) = \mathbb{E}_{y\sim\pi}[r(y)], \qquad J_\lambda(\pi) = J_0(\pi) + \lambda\,\mathbb{E}_{y,y'\sim\pi}[d(y,y')],$$

where $\lambda > 0$ controls the strength of variation.

The key observation is that J_0 is indifferent once all probability lies on correct solutions: collapsing onto one correct mode and covering many can achieve the same objective value. A small positive λ can break this tie without overpowering correctness.

Theorem 1 (Variation promotes mode coverage)**.** *Suppose correct trajectories form $M \geq 2$ reasoning modes with greater cross-mode than within-mode dissimilarity. Let $D_{\max} = \max_{y,y'} d(y,y')$. If*

$$0 < \lambda < \frac{1}{2D_{\max}},$$

then every global maximizer of J_λ still places all probability on correct trajectories, but assigns positive probability to every correct mode:

$$w_m(\pi) = \sum_{y\in G_m} \pi(y) > 0, \qquad m = 1,\dots,M.$$

In contrast, J_0 admits optima concentrated on a single correct mode. Under symmetric modes, J_λ further selects the uniform distribution over all correct trajectories.

Thus, the simplified analysis captures the principle behind our reward design: once selection favors good solutions, variation breaks ties among them in favor of broader mode coverage. Full assumptions and proofs are in Appendix D.

4 Experiments

4.1 Experimental Setup

Training and evaluation. We train on the standard **MATH-TRAIN** set (Hendrycks et al., 2021) and, following TTRL (Zuo et al., 2025), on the smaller **MATH-500** and **AIME24** sets (Li et al., 2024). All training is fully label-free: only problem statements are used, without reference answers or solutions. We evaluate on **MATH-500**, **AIME24**, **AIME25**, **AMC** (Li et al., 2024), and **GPQA-Diamond (GPQA)** (Rein et al., 2024). When evaluation uses the same prompts seen during training, we treat the result as *transductive adaptation*; performance on unseen benchmarks measures held-out transfer. Our main experiments compare Evol-RL with TTRL across training distributions and model scales, while broader label-free methods are compared under matched representative protocols in Section 4.3. Full training details are provided in Appendix A; additional architecture, efficiency, similarity-validation, and sensitivity results are reported in Appendix B.

4.2 Main Results

Table 1 compares Evol-RL with TTRL across two model scales and three label-free training sets. Averaged over all model–training– evaluation combinations, Evol-RL improves pass@1 by 5.3 points and pass@16 by 11.8 points over TTRL. The gains are particularly large on harder benchmarks: Qwen3-4B trained on MATH-TRAIN gains 24.4 points in AIME24 pass@16, while training on AIME24 improves held-out AIME25 pass@16 by 23.5 points. The substantially larger gain in pass@16 is consistent with the intended effect of preserving a broader set of solutions without sacrificing their overall accuracy.

The improvements remain consistent across Qwen3-4B/8B and across MATH-TRAIN, MATH-500, and AIME24, indicating that the benefit is not tied to a particular model scale or training distribution. For example, Qwen3-4B trained on MATH-500 improves AIME24 pass@16 from 28.0 to 43.2 and AIME25 from 29.9 to 41.9, with similar trends for the 8B model.

Table 1: TTRL and EVOL-RL across training settings. Each cell reports pass@1/pass@16 averaged over 32 rollouts; Δ denotes EVOL-RL minus 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

The same pattern holds on benchmarks that are always unseen during training. Across the six training configurations, the average gains on AIME25 are +7.8 points in pass@1 and +12.2 in pass@16; on GPQA, they are +3.2 and +11.6 points, respectively. These results indicate that balancing selection with variation improves not only adaptation to the training distribution but also transfer to held-out reasoning tasks.

4.3 COMPARISON WITH OTHER LABEL-FREE METHODS

To assess whether the gains extend beyond TTRL, we compare EVOL-RL with several label-free self-improvement methods under a matched MATH-500 protocol using Qwen3-4B-Base: TTRL (Zuo et al., 2025), EMPO (Zhang et al., 2025b), Co-Reward (Zhang et al., 2025c), and the recent ECHO (Zhao et al., 2026). All methods use the same training data, rollout budget, optimization settings, and evaluation protocol, so we treat these results as matched-protocol comparisons rather than exact reproductions of independently tuned pipelines.

Table 2: Matched comparison on Qwen3-4B-Base trained on MATH-500. Each cell reports pass@1/pass@16.

Method	MATH	AIME24	AIME25
Base Model	67.4/89.6	10.0/32.4	5.5/30.0
TTRL	79.3/83.2	10.0/28.0	7.2/29.9
EMPO	**79.9**/88.9	11.8/31.6	9.5/26.3
Co-Reward	77.6/89.2	13.6/25.9	10.5/30.3
ECHO	78.5/86.4	10.6/30.6	10.5/31.7
EVOL-RL	79.8/**93.8**	**19.0/43.2**	**16.1/41.9**

As shown in Table 2, this tension is not unique to TTRL. Several label-free baselines improve pass@1 while failing to preserve pass@16, particularly on AIME24 and AIME25. In contrast, EVOL-RL improves both metrics across all three benchmarks, suggesting that explicitly preserving variation helps avoid the loss of solution coverage during self-improvement. Additional matched comparisons with EM-RL-Token/Sequence (Agarwal et al., 2025) and Self-Consistency (Wang et al., 2022; Huang et al., 2023) are reported in Appendix B.2.

Table 3: Performance of Qwen3-4B-Base with EVOL-RL and its ablations. 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

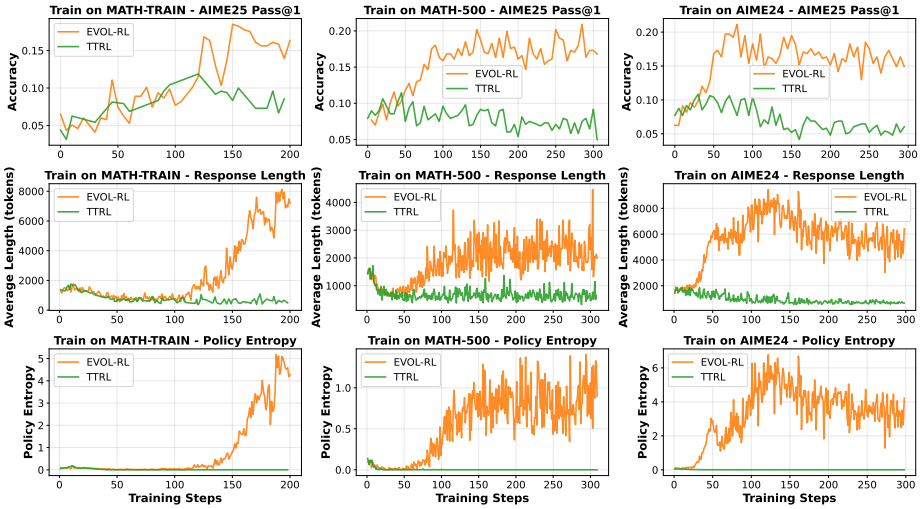

Figure 3: Training dynamics of TTRL and EVOL-RL. Columns correspond to training on MATH-TRAIN, MATH-500, and AIME24; rows show held-out AIME25 pass@1, response length, and policy entropy.

4.4 ABLATION STUDY

Setup. We ablate the three components of EVOL-RL: novelty-based within-group credit allocation, entropy regularization, and asymmetric clipping. We remove each component individually, and remove the two supporting mechanisms jointly, on Qwen3-4B-Base. Results are shown in Table 3.

Component roles depend on the training distribution. With MATH-500 training, removing novelty causes the largest degradation in pass@16 on harder held-out tasks, reducing AIME24 from 43.2 to 27.0 and AIME25 from 41.9 to 34.8. This indicates that within-group novelty is particularly important when majority-based selection alone provides little pressure to preserve alternative reasoning paths.

The pattern changes with AIME24 training. Here, removing entropy regularization or asymmetric clipping causes larger drops than removing novelty on several metrics. This is consistent with their supporting roles: entropy regularization maintains a supply of varied rollouts, while asymmetric clipping preserves strong updates from high-advantage samples. Overall, the ablation suggests that the three components are complementary, with their relative importance depending on the amount of variation already present in the training distribution.

4.5 TRAINING DYNAMICS: THE EVOLVING POINT AND RECOVERY

To understand how EVOL-RL changes the training trajectory, we compare its dynamics with TTRL in Figure 3. Results for Qwen3-8B-Base show the same overall pattern and are reported in Appendix B.6.

Stage 1: Initial Collapse Under Majority Signal. Across all three training sets, both methods initially show sharp decreases in policy entropy and average response length. Their trajectories remain similar during this phase, indicating that the majority-based selection signal initially dominates the update and concentrates the policy. TTRL remains in this low-entropy, short-response regime throughout the rest of training.

Stage 2: The Evolving Point and Coordinated Recovery. EVOL-RL subsequently diverges from TTRL at a clear *evolving point*, whose timing varies across training sets. After this point, policy entropy and response length recover together with held-out AIME25 pass@1, eventually stabilizing at higher levels, while TTRL remains largely flat. Pass@16 trajectories show the same divergence (Appendix B.7). The evolving point therefore marks the transition from the initial selection-dominated phase to a recovery phase in which broader solution coverage re-emerges.

This recovery is consistent with the complementary roles identified in the ablation study: entropy regularization maintains candidate diversity, novelty redistributes credit toward less redundant trajectories within each selection group, and asymmetric clipping preserves stronger positive-advantage updates.

4.6 WHAT DOES VARIATION PRESERVE?

Corrective alternatives. We first test whether variation keeps a correct answer reachable when the current consensus is wrong. On Qwen3-4B-Base trained on MATH-500, we compare step 40, before TTRL and EVOL-RL begin to diverge, with step 105, after the recovery emerges. We measure *correct-minority availability*: among prompts whose majority answer is incorrect, the fraction with at least one correct minority response. As shown in Table 4, this availability falls from 45.1% to 38.3% under TTRL, but rises from 45.3% to 55.5% under EVOL-RL, together with higher pass@16. Thus, preserving variation helps keep corrective solutions reachable when the current consensus is wrong. Full details are provided in Appendix B.3.

Table 4: Training dynamics from step 40 to 105 on Qwen3-4B-Base with MATH-500.

	TTRL	EVOL-RL
Correct-minority	45.1→38.3	45.3→**55.5**
MATH pass@16	89.0→88.5	89.8→**93.1**

Distinct successful reasoning strategies. We next ask whether the preserved variation corresponds to different successful approaches rather than merely diverse failures. On Qwen3-4B-Base trained on AIME24, we evaluate three checkpoints—the base model, one before the evolving point, and the final model—using 16 rollouts on the same 30 AIME24 problems. Restricting to pairs in which both responses are correct, Qwen3-235B-A22B-Instruct judges whether the two solutions use the same or different mathematical approaches. The different-approach rate increases from 11.1% in the base model to 22.2% in the final model, while the number of comparable correct-solution pairs grows from 25 to 51. This complements the correct-minority analysis: EVOL-RL preserves not only corrective alternatives when consensus is wrong, but also multiple successful reasoning paths. Full details are provided in Appendix B.4.

4.7 EVOL-RL COMPONENTS ALSO STRENGTHEN SUPERVISED GRPO (RLVR)

Setup. We apply the three components of EVOL-RL to a standard supervised GRPO baseline trained on the MATH training set (Hendrycks et al., 2021) with a ground-truth verifier for two epochs. Unlike our main label-free setting, selection here is determined directly by answer correctness rather than a self-generated majority signal. Figure 4 reports the results.

The full combination, `GRPO+ClipHigh+Ent+Novelty`, achieves the strongest overall performance. Compared with standard GRPO, it improves pass@16 by 12.7 points on AIME24 and 9.0 points on AIME25, while also improving pass@1 by 8.5 and 7.1 points, respectively. The same configuration improves pass@16 on MATH, AMC, and GPQA by 3.5, 4.8, and 7.1 points, respectively, while improving or essentially maintaining pass@1. These results show that the benefit of preserving variation is not limited to label-free majority-based training: even when a ground-truth verifier provides a reliable selection signal, maintaining diverse candidates and differentiating among correct reasoning trajectories can further improve reasoning performance.

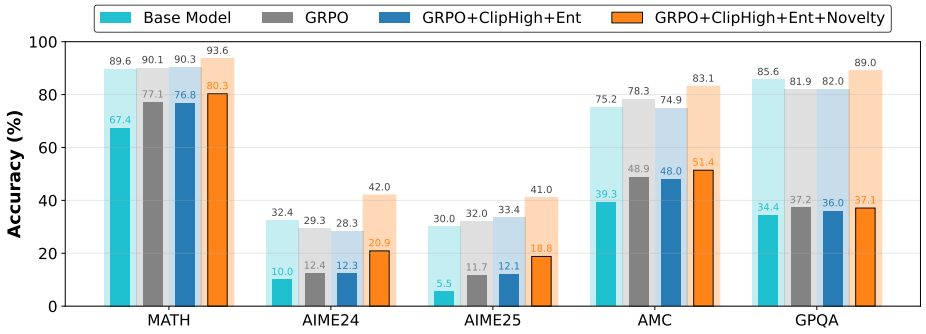

Figure 4: Performance of EVOL-RL components applied to supervised GRPO. Qwen3-4B-Base is trained on MATH with a ground-truth verifier. Outer bars show pass@16 and inner bars show pass@1.

Table 5: Transfer performance of the Qwen3-8B-Base model on broader reasoning benchmarks after label-free training on MATH-TRAIN.

Model	**MMLU-Pro**		**SuperGPQA**		**BBEH**	
	Pass@1	Pass@4	Pass@1	Pass@4	Pass@1	Pass@4
Qwen3-8B-Base	47.3	74.5	26.5	54.1	10.4	24.0
w/TTRL	53.4	73.9	29.7	53.3	**12.1**	24.1
w/EVOL-RL	**55.3**	**78.5**	**30.2**	**57.0**	11.5	**24.9**

4.8 TRANSFER TO BROADER REASONING BENCHMARKS

To assess whether the improvements learned from mathematical data transfer to broader reasoning tasks, we evaluate Qwen3-8B-Base after label-free training on MATH-TRAIN on **MMLU-Pro** (Wang et al., 2024), **SuperGPQA** (Team et al., 2025), and **BBEH** (Kazemi et al., 2025). None of these benchmarks is used during training. The results are reported in Table 5.

A similar pass@1–pass@k tension appears beyond the mathematical evaluation sets. TTRL improves pass@1 substantially over the base model, but its pass@4 falls slightly below the base model on MMLU-Pro (73.9 vs. 74.5) and SuperGPQA (53.3 vs. 54.1), indicating that stronger single-response performance does not always translate into broader coverage. In contrast, EVOL-RL improves pass@4 over both TTRL and the base model on all three benchmarks: 78.5 on MMLU-Pro, 57.0 on SuperGPQA, and 24.9 on BBEH. Pass@1 also improves over TTRL on MMLU-Pro and SuperGPQA, while remaining slightly lower on BBEH (11.5 vs. 12.1). These results extend the pattern observed in our main experiments: preserving variation during label-free mathematical training transfers positively to broader held-out reasoning tasks, with the clearest gains appearing when multiple attempts are considered.

5 CONCLUSION

In this work, we identify *selection-induced diversity collapse* as a failure mode of label-free self-improvement: when supervision is repeatedly constructed from a model's own outputs, selection can progressively narrow the candidate pool and make alternative correct solutions harder to recover. We propose EVOL-RL, which addresses this problem by balancing *selection* and *variation*: majority voting provides a stable selection signal, while novelty-based credit allocation preserves variation within each selection group, supported by entropy regularization and asymmetric clipping. Across training settings, model scales, and matched self-improvement baselines, EVOL-RL improves reasoning performance while better preserving solution coverage. Training analyses further show that it retains correct alternatives when the current consensus is wrong, supporting the role of variation in sustaining future opportunities for correction.

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.

Mucong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, and Furong Huang. Sail: Self-improving efficient online alignment of large language models. *arXiv preprint arXiv:2406.15567*, 2024.

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

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. Serl: Self-play reinforcement learning for large language models with limited data. *arXiv preprint arXiv:2505.20347*, 2025.

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.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 1051–1068, 2023.

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.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard, 2025. URL https://arxiv.org/abs/2502.19187.

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.

Sangkyu Lee, Sungdong Kim, Ashkan Yousefpour, Minjoon Seo, Kang Min Yoo, and Youngjae Yu. Aligning large language models by on-policy self-judgment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11442–11459, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.617. URL https://aclanthology.org/2024.acl-long.617/.

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.

Wendi Li, Shawn Im, and Sharon Li. Cyclical entropy eruption: Entropy dynamics in agent reinforcement learning. *arXiv preprint arXiv:2605.27954*, 2026.

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.

Xun Liang, Shichao Song, Zifan Zheng, Hanyu Wang, Qingchen Yu, Xunkai Li, Rong-Hua Li, Yi Wang, Zhonghao Wang, Feiyu Xiong, and Zhiyu Li. Internal consistency and self-feedback in large language models: A survey, 2024. URL https://arxiv.org/abs/2407.14507.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. Clue: Non-parametric verification from experience via hidden-state clustering. *arXiv preprint arXiv:2510.01591*, 2025a.

Zhenwen Liang, Sidi Lu, Wenhao Yu, Kishan Panaganti, Yujun Zhou, Haitao Mi, and Dong Yu. Can llms guide their own exploration? gradient-guided reinforcement learning for llm reasoning. *arXiv preprint arXiv:2512.15687*, 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*, 2025a.

Rui Liu, Dian Yu, Lei Ke, Haolin Liu, Yujun Zhou, Zhenwen Liang, Haitao Mi, Pratap Tokekar, and Dong Yu. Stable and efficient single-rollout rl for multimodal reasoning. *arXiv preprint arXiv:2512.18215*, 2025b.

Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, XiongHui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. Language model self-improvement by reinforcement learning contemplation. In *International Conference on Learning Representations*, volume 2024, pp. 48842–48868, 2024.

Jaesung R Park, Junsu Kim, Gyeongman Kim, Jinyoung Jo, Sean Choi, Jaewoong Cho, and Ernest K Ryu. Clip-low increases entropy and clip-high decreases entropy in reinforcement learning of large language models. *arXiv preprint arXiv:2509.26114*, 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.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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.

P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

Thinking Machines Lab. Announcing Tinker, October 2025. URL https://thinkingmachines.ai/news/announcing-tinker/.

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*, 2025a.

Xiangqi Wang, Yue Huang, Yanbo Wang, Xiaonan Luo, Kehan Guo, Yujun Zhou, and Xiangliang Zhang. Adareasoner: Adaptive reasoning enables more flexible thinking. *arXiv preprint arXiv:2505.17312*, 2025b.

Xiangqi Wang, Yue Huang, Yujun Zhou, Xiaonan Luo, Kehan Guo, and Xiangliang Zhang. Causally-enhanced reinforcement policy optimization. *arXiv preprint arXiv:2509.23095*, 2025c.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025d.

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.

Kongcheng Zhang, Qi Yao, Shunyu Liu, Yingjie Wang, Baisheng Lai, Jieping Ye, Mingli Song, and Dacheng Tao. Consistent paths lead to truth: Self-rewarding reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.08745*, 2025a.

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*, 2025b.

Zizhuo Zhang, Jianing Zhu, Xinmu Ge, Zihua Zhao, Zhanke Zhou, Xuan Li, Xiao Feng, Jiangchao Yao, and Bo Han. Co-reward: Self-supervised reinforcement learning for large language model reasoning via contrastive agreement. *arXiv e-prints*, pp. arXiv–2508, 2025c.

Chu Zhao, Enneng Yang, Yuting Liu, Jianzhe Zhao, and Guibing Guo. Echo: Entropy-confidence hybrid optimization for test-time reinforcement learning. *arXiv preprint arXiv:2602.02150*, 2026.

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

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

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*, 2025c.

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.

Haomin Zhuang, Yujun Zhou, Taicheng Guo, Yue Huang, Fangxu Liu, Kai Song, and Xiangliang Zhang. Exploring multi-temperature strategies for token-and rollout-level control in rlvr. *arXiv preprint arXiv:2510.08892*, 2025.

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 Training Configuration.

Our main experiments use two open-source base models: **Qwen3-4B-Base** and **Qwen3-8B-Base**. Additional architecture and scale experiments with OctoThinker-8B-Hybrid-Base and Qwen3.5-35B-A3B-Base are reported in Appendix B.1. Our main 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).

A.2 System Prompt

For the main 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.3 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.4 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-Embedding-4B** 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.5 Hyperparameter Settings

For our main local label-free experiments, we largely follow the settings established by TTRL to ensure a fair comparison. The general hyperparameters are detailed in Table 6, and the settings specific to our EVOL-RL method are listed in Table 7.

Table 6: 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 7: 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

A.6 Computational Resources

The main local experiments were conducted on a server with 8 NVIDIA H20 GPUs. The Qwen3.5-35B-A3B-Base scale experiment was run through Tinker using LoRA; its configuration and results are reported in Appendix B.1.

B Additional Experimental Results

B.1 Effectiveness Across Model Architectures and Scale

We further examine whether the gains of EVOL-RL extend beyond the Qwen3-4B/8B configurations used in our main experiments along two complementary dimensions: model architecture and model scale. We first evaluate a different 8B architecture, OctoThinker-8B-Hybrid-Base, and then conduct a short-horizon scale check on Qwen3.5-35B-A3B-Base.

OctoThinker-8B-Hybrid-Base. To test whether the observed behavior is specific to the Qwen3 architecture, we apply EVOL-RL and TTRL to **OctoThinker-8B-Hybrid-Base** (Wang et al., 2025d), using the same label-free training setup on MATH-500. As shown in Table 8, TTRL substantially improves MATH pass@1 (from 33.8% to 63.8%), but its pass@16 decreases relative to the base model on AIME24 (13.5% to 10.8%) and GPQA (85.7% to 71.5%). In contrast, EVOL-RL achieves comparable MATH pass@1 while improving pass@16 over TTRL across all five benchmarks, including gains of 19.5 points on AIME24 and 11.4 points on AIME25.

Table 8: Comparison of models trained with TTRL and EVOL-RL on MATH-500 on OctoThinker-8B-Hybrid-Base. Each cell shows pass@1/pass@16 (averaged over 32 rollouts). Δ uses red (+) for positive and blue for negative values.

Model	MATH	AIME24	AIME25	AMC	GPQA
OctoThinker-8B-Hybrid-Base					
Base Model	33.8/79.8	1.5/13.5	1.3/13.9	16.2/56.9	26.3/85.7
w/TTRL	63.8/76.4	2.8/10.8	2.1/11.0	27.9/54.7	31.9/71.5
w/EVOL-RL	63.2/86.3	9.0/30.3	7.2/22.4	34.1/65.6	33.2/85.7
Δ	-0.6/+9.9	+6.2/+19.5	+5.1/+11.4	+6.2/+10.9	+1.3/+14.2

Qwen3.5-35B-A3B-Base. We next examine whether the same pattern extends beyond the model sizes used in our main experiments. We train Qwen3.5-35B-A3B-Base (Qwen Team, 2026) on AIME24 using LoRA through Tinker (Thinking Machines Lab, 2025), with both TTRL and EVOL-RL trained for 30 steps. Because of this limited training horizon, we treat the experiment as a scale check rather than a fully converged comparison. As shown in Table 9, both methods improve over the base model on AIME24, but EVOL-RL achieves higher pass@1 and pass@16 than TTRL on both AIME24 and held-out AIME25. The difference is again clearer in multi-sample coverage, with pass@16 improving from 86.2 to 88.4 on AIME24 and from 67.9 to 71.4 on AIME25.

Table 9: Short-horizon comparison of TTRL and EVOL-RL on Qwen3.5-35B-A3B-Base trained on AIME24 for 30 steps using LoRA. Each cell shows pass@1/pass@16. Δ uses red (+) for positive and blue for negative values.

Model	AIME24	AIME25
Qwen3.5-35B-A3B-Base		
Base Model	57.0/80.8	44.9/67.6
w/TTRL	63.0/86.2	45.3/67.9
w/EVOL-RL	63.6/88.4	48.2/71.4
Δ	+0.6/+2.2	+2.9/+3.5

Together, these results provide complementary evidence that the improvements of EVOL-RL are not confined to the particular architecture or model sizes used in our main Qwen3-4B/8B experiments.

B.2 ADDITIONAL SELF-IMPROVEMENT BASELINES

The main text compares EVOL-RL with TTRL, EMPO, Co-Reward, and ECHO under a matched MATH-500 protocol. Here, we provide additional matched comparisons on MATH-TRAIN against Self-Consistency (Wang et al., 2022; Huang et al., 2023) and the confidence-based EM-RL-Token and EM-RL-Sequence objectives (Agarwal et al., 2025). All methods use Qwen3-4B-Base and the same label-free training and evaluation protocol. Results are reported in Table 10.

Baseline implementation. We use the official TTRL implementation with the shared backbone and a maximum response length of 12,288 tokens. EMPO retains the semantic-entropy reward from its official verl implementation, and Co-Reward retains its defining contrastive-agreement reward within the shared training framework. For ECHO, we follow the authors' released MATH reproduction configuration, using the official policy-update components and their method-specific default hyperparameters. These comparisons share the backbone, training prompts, rollout budget, and evaluation protocol, while preserving method-specific reward and update settings; they are not reproductions of each independently tuned pipeline.

A similar pattern emerges. TTRL, EM-RL, and Self-Consistency improve pass@1 on several benchmarks, but their pass@16 gains are less consistent. On AIME24, for example, all four alternatives remain below the base model in pass@16 (23.2–31.4 vs. 32.4), whereas EVOL-RL reaches 47.6. On AIME25, EVOL-RL achieves 17.5/39.9 in pass@1/pass@16, compared with 10.5/30.8 for EM-RL-

Table 10: Extended comparison of baseline methods on Qwen3-4B-Base, trained in a label-free setting using MATH-Train. Each cell shows pass@1/pass@16, and the highest value in each column is bolded.

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
w/TTRL	75.4/86.9	12.1/23.2	6.8/28.6	42.5/75.2	36.5/81.4
w/EM-RL-Token	76.0/90.6	12.5/31.3	10.5/30.8	46.6/77.7	36.8/82.6
w/EM-RL-Sequence	67.4/89.9	10.6/31.4	7.1/28.8	39.6/73.7	34.5/86.2
w/Self-Consistency	76.0/89.7	12.5/30.4	10.4/33.6	48.1/78.1	35.9/81.2
w/EVOL-RL	**80.0/93.3**	**20.7/47.6**	**17.5/39.9**	**51.4/80.3**	**37.2/88.7**

Token and 10.4/33.6 for Self-Consistency. These additional comparisons reinforce the trend observed in the main text: improving the self-generated training signal does not necessarily preserve solution coverage, while explicitly maintaining variation improves both accuracy and coverage.

B.3 CORRECT-MINORITY AVAILABILITY

We further examine whether correct alternatives remain reachable when the current majority answer is wrong. At checkpoint t, we define *correct-minority availability* as

$$\frac{\#\{\text{prompts with an incorrect majority and at least one correct minority response}\}}{\#\{\text{prompts with an incorrect majority}\}}.$$

The denominator is recomputed independently at each checkpoint.

We analyze Qwen3-4B-Base trained on MATH-500. Step 40 is chosen as a pre-divergence checkpoint: TTRL and EVOL-RL still have nearly identical entropy and similar accuracy and correct-minority availability (Figure 3). By step 105, their training trajectories have clearly diverged.

The two methods exhibit different trajectories. Under TTRL, pass@1 improves slightly, but correct-minority availability, pass@16, entropy, and response length all decrease. Thus, majority-only training can strengthen the current prediction while making corrective alternatives less likely to remain sampled. Under EVOL-RL, correct-minority availability instead rises together with pass@16, entropy, and response length.

The same pattern persists over the full run. From step 0 to step 180, correct-minority availability decreases from 46.6% to 34.4% under TTRL, while remaining roughly stable under EVOL-RL (50.1% to 50.9%). Over the same interval, MATH pass@16 changes from 88.3% to 88.0% for TTRL and from 89.3% to 93.9% for EVOL-RL. These results indicate that EVOL-RL preserves access to correct alternatives when the self-generated majority is wrong, although it does not guarantee that an incorrect majority will eventually be overturned.

Table 11: Training dynamics from step 40 to 105 on Qwen3-4B-Base trained on MATH-500.

Metric	TTRL	EVOL-RL
MATH pass@1	75.4 → 77.8	76.1 → 78.6
MATH pass@16	89.0 → 88.5	89.8 → 93.1
Correct-minority availability	45.1 → 38.3	45.3 → 55.5
Mean response length	524 → 488	593 → 1,983

B.4 DISTINCT SUCCESSFUL REASONING STRATEGIES

We further examine whether the variation preserved by EVOL-RL corresponds to genuinely different successful reasoning strategies, rather than merely diverse outputs or failure modes. We analyze Qwen3-4B-Base trained on AIME24 at three checkpoints: the base model, one checkpoint before the evolving point, and the final model. At each checkpoint, we generate 16 rollouts for the same 30 AIME24 problems.

We restrict the analysis to pairs in which both solutions are correct. This avoids conflating reasoning diversity with diversity among incorrect solutions: at the base checkpoint, most generated responses are incorrect, so evaluating all pairs would largely measure different failure modes rather than variation among successful reasoning strategies.

For each eligible pair, Qwen3-235B-A22B-Instruct judges whether the two solutions use the same or different mathematical approaches. The judge receives only the problem and two anonymized solutions, labeled Solution A and Solution B, without information about their checkpoint of origin. Each pair is judged twice with the solution order reversed, and we retain only pairs for which the two judgments agree. Results are reported in Table 12.

Table 12: Reasoning-strategy diversity across training checkpoints on AIME24. The analysis includes only pairs in which both solutions are correct.

Checkpoint	Different-Approach Rate	Correct Pairs
Base	11.1%	25
Before evolving point	14.3%	45
Final	22.2%	51

Among correct solutions, the fraction judged to use different mathematical approaches increases from 11.1% for the base model to 22.2% for the final model. The number of comparable correct-solution pairs also increases from 25 to 51 as accuracy on these problems rises from 8.5% to 18.8%. Thus, the final model not only produces more correct solutions, but its correct solutions are also more likely to represent distinct reasoning approaches.

Because the number of correct-solution pairs is limited, particularly for the base checkpoint, we treat this analysis as direct but exploratory evidence. Together with the correct-minority analysis in Appendix B.3, it suggests that EVOL-RL preserves both corrective alternatives when the current consensus is wrong and multiple successful reasoning paths when correct solutions are available.

B.5 ANALYSIS OF THE MAJORITY-VOTE SIGNAL

To examine how preserving variation affects the self-generated supervision available later in training, we track the accuracy of the majority-voted answer over time. At each checkpoint, we derive a consensus answer from 16 rollouts and evaluate it against the ground-truth answer on AIME25; ground truth is used only for this analysis and is never provided during label-free training. Figure 5 reports the resulting maj@16 trajectories.

Across all three training distributions, TTRL and EVOL-RL initially follow similar trajectories. TTRL's majority-vote accuracy improves early but then largely plateaus. In contrast, EVOL-RL subsequently diverges near the evolving point and continues to improve its majority-vote accuracy, eventually stabilizing at a higher level.

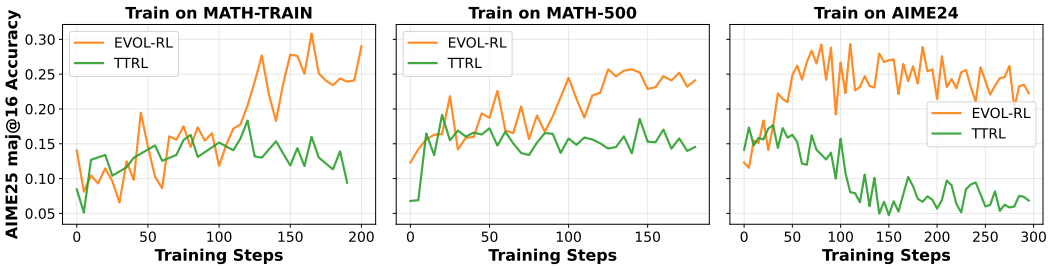

Figure 5: Majority-vote accuracy (maj@16) on AIME25 throughout training. The models are trained on (**Left**) MATH-TRAIN, (**Middle**) MATH-500, and (**Right**) AIME24.

Because the majority-voted answer is itself used to construct the training signal, this difference is important for the self-improvement loop. Preserving variation does not merely maintain alternative responses at the current step; it also leaves a broader candidate pool from which later majority signals are constructed. The continued increase in maj@16 under EVOL-RL is therefore consistent with progressively improving the quality of its future self-generated supervision, whereas TTRL's supervision quality changes little after its early improvement.

B.6 Training Dynamics of 8B Models

We further examine whether the training dynamics observed with Qwen3-4B-Base persist at a larger model scale. Figure 6 reports AIME25 pass@1, average response length, and policy entropy throughout training for Qwen3-8B-Base across MATH-TRAIN, MATH-500, and AIME24.

Across all three training distributions, the 8B model exhibits the same two-stage pattern observed in the 4B experiments. During the initial stage, TTRL and EVOL-RL follow similar trajectories: policy entropy and response length decrease as majority-based selection concentrates the policy. TTRL largely remains in this concentrated regime for the rest of training.

EVOL-RL, however, subsequently diverges at an evolving point. After this point, policy entropy and response length recover together with held-out AIME25 pass@1, while the corresponding TTRL trajectories remain substantially flatter. The timing of the evolving point differs across training distributions, but the qualitative pattern is consistent. These results show that the recovery dynamics observed in the main 4B experiments are not specific to a single model scale.

B.7 Pass@16 Training Curves

To complement the training dynamics in Section 4.5, Figure 7 reports the pass@16 trajectories on AIME24 and AIME25 across all three training settings. TTRL's pass@16 generally declines or plateaus over training, whereas EVOL-RL's pass@16 recovers after the evolving point and stabilizes at a higher level. This provides additional evidence that the recovery observed in policy entropy and response length is accompanied by broader solution coverage.

B.8 Analysis of the Computational Overhead from the Novelty Reward

A practical consideration for our method is the additional computational load introduced by the embedding-based novelty reward. To rigorously quantify this, we tracked the wall-clock time per training step, decomposing it into two parts: (1) the Novelty Reward Calculation (comprising $B \times N$ embedding model calls and the $O(BN^2)$ similarity matrix computation, where B is the batch size and N is the number of rollouts), and (2) the Total Training Process (including rollout generation, reward calculation and model updates). The results are plotted in Figure 8.

The analysis reveals two key insights:

- **Scaling with Reasoning Length:** As expected, both curves rise over time. This correlates with our findings in Figure 3 that the model learns to generate significantly longer CoT during training. Longer responses require more time to generate (Training) and more time to embed (Reward Calculation).

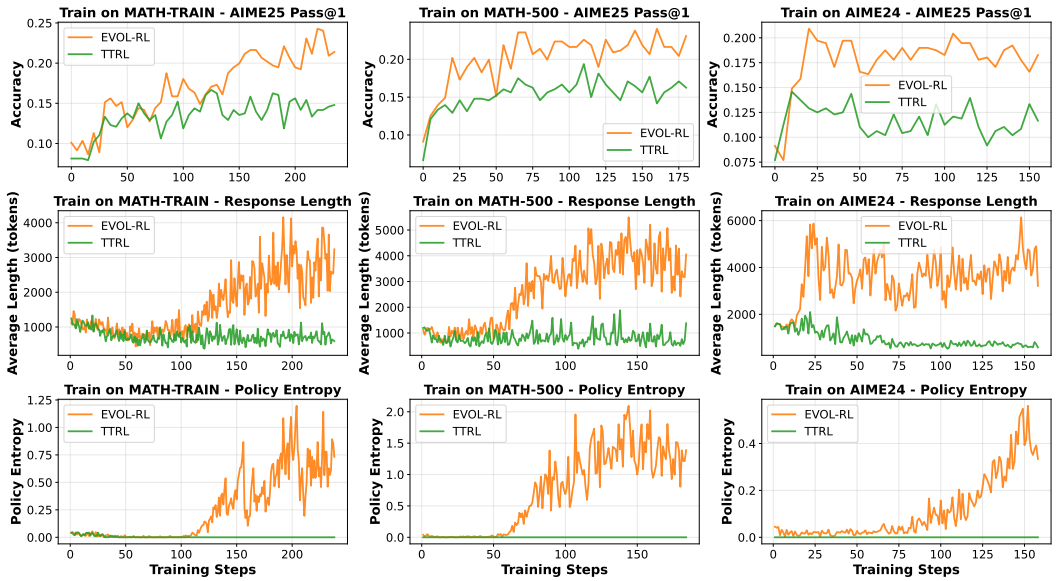

Figure 6: Training dynamics of TTRL and EVOL-RL on Qwen3-8B-Base. Columns correspond to training on MATH-TRAIN, MATH-500, and AIME24; rows show held-out AIME25 pass@1, average response length, and policy entropy.

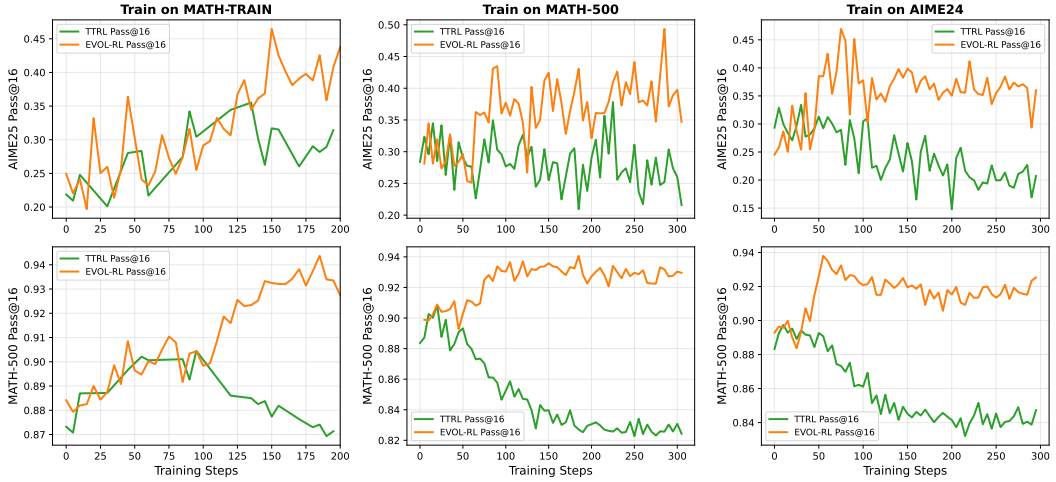

Figure 7: Pass@16 training curves on AIME24 and AIME25 across three training settings (MATH-TRAIN / MATH-500 / AIME24). TTRL's pass@16 consistently declines, while EVOL-RL's steadily recovers and grows after the evolving point.

- **Low Relative Overhead:** while the reward cost scales with response length (requiring embedding of longer sequences), the Novelty Reward Calculation (red curve) remains a small portion of the overall runtime, stabilizing around 100 seconds.

The occasional sharp spikes in the Red curve are anomalous and likely attributable to external API connection instabilities. We conclude that the embedding-based reward introduces a manageable overhead relative to the overall cost of training the reasoning model.

B.9 CASE STUDY: VALIDITY OF THE SEMANTIC NOVELTY PROXY

A potential concern regarding our method is whether semantic similarity in an embedding space is a meaningful proxy for genuine reasoning novelty. To validate this, we conducted a case study, shown

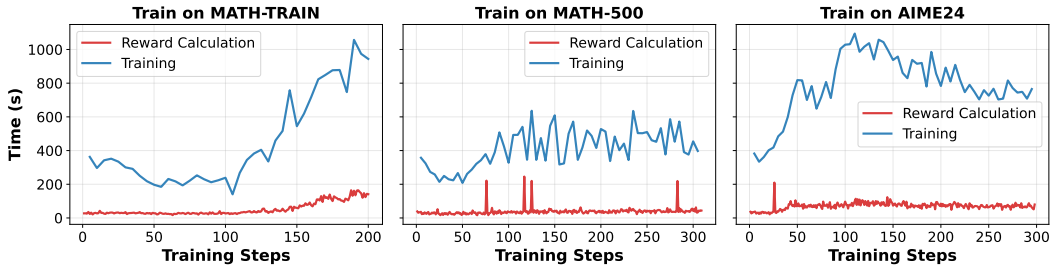

Figure 8: Wall-clock time (seconds) per training step for **Novelty Reward Calculation** (Red) versus the **Total Training Process** (Blue) across three datasets. While both costs naturally increase as the model evolves to generate longer reasoning paths, the reward calculation overhead remains a minor fraction of the total computational load.

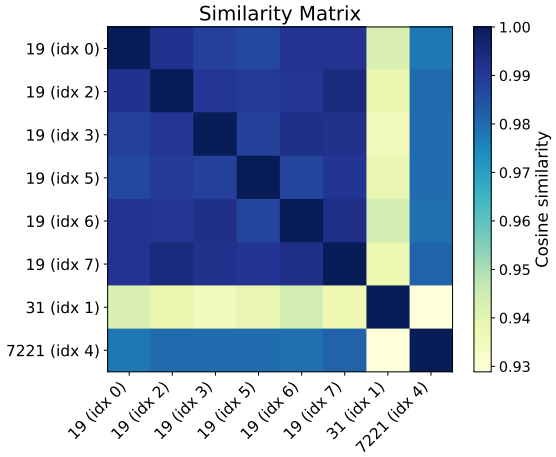

Figure 9: Case study of the reasoning path cosine similarity matrix for 8 rollouts on a single AIME25 problem. The axes are labeled with the `final_answer(index_i)` for each rollout. The 8 rollouts produced three distinct final answers: "19" (6 times), "31" (1 time), and "7221" (1 time).

in Figure 9, by generating 8 rollouts for a single AIME25 problem using Qwen3-4B-Instruct-2407 and computing their reasoning path similarity matrix with Qwen3-Embedding-4B.

The results provide strong empirical evidence for the validity of this proxy. As shown in the figure, the matrix reveals distinct, high-similarity clusters that align with the reasoning logic. Specifically, the 6 rollouts that produced the same final answer ("19") have an extremely high intra-cluster similarity (the 6x6 block at the top-left). Crucially, these paths are semantically distinct from the path that answered "31" and the path that answered "7221", showing clear separation in the embedding space.

This demonstrates that the embedding space does successfully capture and cluster the core logic of the reasoning paths, distinguishing between different reasoning trajectories.

B.10 LLM-Judge Validation of the Cosine Similarity Proxy

We validate that cosine similarity in the embedding space tracks semantic reasoning differences rather than surface artifacts such as length, formatting, or phrasing. To this end, we ask an independent LLM judge—DeepSeek-V3.2, drawn from a different model family than the embedding—to label whether each pair of rollouts shares the same mathematical approach. The evaluation pool consists of 60 AIME questions with 64 EVOL-RL rollouts each (Qwen3-4B-Base, MATH-500 training); for each question we sample pairs from the top-third (high-similarity) and bottom-third (low-similarity) of its cosine distribution and submit them anonymously to the judge.

Table 13: LLM-judge agreement with cosine-similarity ranking, grouped by response length.

Length	# Q	High-sim→SAME	Low-sim→DIFF	Agreement	Sim. Gap
$<$ 3K	12	66.7%	83.3%	75.0%	+0.301
3K–4.5K	20	80.0%	91.7%	85.8%	+0.466
4.5K–6K	13	69.2%	89.7%	79.5%	+0.378
6K+	15	88.9%	95.6%	92.2%	+0.550
Total	**60**	**77.2%**	**90.6%**	**83.9%**	**+0.436**

Table 14: Surface-rewriting robustness on 300 solutions (30 AIME24/25 questions $\times$ 5 each). Refined solutions remain close to their originals, while cross-solution pairs are clearly more distant.

Comparison	Cosine Similarity
Refined vs. original (within-pair)	0.834
Cross-solution (different reasoning)	0.681
Gap	**+0.154**

Across the 60 questions, the judge agrees with the cosine ranking on 83.9% of sampled pairs. The agreement is asymmetric in a way that supports the design: high-similarity pairs are labeled SAME at 77.2%, while low-similarity pairs are labeled DIFFERENT at 90.6%—i.e., the embedding rarely flags two semantically distinct solutions as similar. The mean cosine gap between SAME and DIFFERENT pairs is +0.436, well above any plausible noise floor.

We further verify that the proxy remains discriminative on long reasoning traces, where all embeddings could in principle be pushed toward a common region. Table 13 bins the questions by response length and shows the opposite of any such collapse: agreement rises with length, from 75.0% on the shortest bin ($<$3K tokens) to 92.2% on the longest ($>$6K tokens), and the SAME/DIFFERENT cosine gap widens correspondingly. Longer traces give the embedding more structural signal to work with, not less.

B.11 Surface-Rewriting Robustness

We next test whether surface-level paraphrasing can inflate the novelty signal: an agent that simply restates the same solution in different language should not be rewarded as "novel" since no real reasoning variation has occurred. We use DeepSeek-V3.2 to surface-rewrite 300 solutions (30 AIME24/25 questions $\times$ 5 solutions each), with explicit instructions to preserve every reasoning step while polishing language and notation.

Across the 300 refine pairs, the cosine similarity between a refined solution and its original is 0.834, well above the average similarity of 0.681 to unrelated solutions on the same question (Table 14). The gap of +0.154 between within-pair and cross-pair similarity is highly significant ($p < 0.001$). Surface rewriting therefore moves a solution only marginally in embedding space, whereas genuinely distinct reasoning paths produce substantially larger displacement. The novelty signal is sensitive to the underlying reasoning structure, not the surface form.

B.12 Cross-Family Embedding Robustness

We finally examine the framework's sensitivity to the embedding model family. Because the policy is Qwen-based, one might wonder whether the Qwen-family embedding contributes inductive biases that another embedding would not, so we re-run the full pipeline after replacing Qwen3-Embedding-4B with OpenAI's text-embedding-3-small—an entirely different model family with a different training corpus and architecture—while keeping every other configuration fixed.

Both embedding choices improve pass@16 over the no-novelty ablation across MATH, AIME24, and AIME25, while maintaining or improving pass@1. On the most challenging benchmark, AIME24 pass@16 reaches 43.2% with the Qwen embedding and 39.7% with the OpenAI embedding—both far above the 27.0% obtained without any novelty signal. The MATH numbers are essentially identical (93.8% with either embedding vs. 88.7% without), and AIME25 follows the same pattern (Table 15).

The 3.5-pp residual gap between the two embeddings on AIME24 pass@16 is consistent with the OpenAI model's 8K-token context limit, which truncates the longer reasoning traces that carry the most distinctive structure. Once that artifact is accounted for, the framework is essentially indifferent to the embedding family.

Table 15: Cross-family embedding ablation; both embeddings improve multi-sample coverage over the no-novelty ablation.

Novelty Signal	MATH p@1/p@16	AIME24 p@1/p@16	AIME25 p@1/p@16
Qwen3-Embedding-4B	79.8/93.8	19.0/43.2	16.1/41.9
OpenAI text-embedding-3-small	79.3/93.8	16.2/39.7	15.6/40.8
No novelty (−Novelty in Table 3)	79.3/88.7	12.1/27.0	11.1/34.8

B.13 Sensitivity to Novelty Strength, Mixing Coefficient α, and Sampling Temperature

We examine sensitivity to three choices that affect how variation is introduced or observed: the strength of novelty-dependent reward variation, the mixing coefficient α used to compute novelty, and the sampling temperature T used during evaluation.

Novelty reward strength. Our default reward uses

$$r_i = \begin{cases} 0.5 + \beta\tilde{u}_i, & y_i = +1, \\ -1 + \beta\tilde{u}_i, & y_i = -1, \end{cases}$$

with $\beta = 0.5$. Thus, β controls how strongly novelty differentiates responses within the same selection group, while $\beta = 0$ removes novelty-dependent ranking but leaves majority selection, entropy regularization, and asymmetric clipping unchanged. We sweep $\beta \in \{0, 0.5, 1.0, 1.5, 2.0\}$ on Qwen3-4B-Base trained with MATH-500. Results are shown in Table 16.

Performance improves substantially from $\beta = 0$ to the default $\beta = 0.5$: averaged across the four benchmarks, pass@16 increases from 58.0 to 67.0. Increasing the strength further does not continue to improve performance. At $\beta = 1.0$, majority responses are still guaranteed to receive higher rewards than minority responses, but performance already begins to decline. At $\beta = 1.5$, the two reward bands meet, and for larger values they overlap, weakening the intended priority of selection over variation. The degradation at large β therefore supports the design choice of using novelty to refine credit within selection groups rather than allowing it to override the selection signal.

Mixing coefficient α. We next vary the coefficient α that balances the mean within-group similarity and maximum-similarity terms in the novelty score. We sweep $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ on Qwen3-4B-Base trained with MATH-500 and report pass@1/pass@16 on the five evaluation benchmarks (Table 17). Performance is essentially flat across the $[0.25, 0.75]$ interior of the range, with pass@1 varying by less than 1 pp and pass@16 by less than 1.5 pp. Even at the extremes ($\alpha = 0$ or $\alpha = 1$), EVOL-RL substantially outperforms TTRL on most metrics, suggesting that the gains do not depend on fine-grained tuning of this coefficient.

Sampling temperature. We finally evaluate the trained EVOL-RL model (Qwen3-4B-Base, MATH-TRAIN) at sampling temperatures $T \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ (Table 18). At $T = 0$, decoding is deterministic, so pass@16 reduces to pass@1. For all nonzero temperatures, multi-sample evaluation provides substantial gains over pass@1, while both metrics remain relatively stable across a broad range of temperatures. Thus, the observed multi-sample gains are not specific to a single evaluation temperature.

Table 16: Sensitivity to novelty reward strength β on Qwen3-4B-Base trained with MATH-500. $\beta = 0.5$ is the default setting. Each cell shows `pass@1`/`pass@16`.

β	MATH	AIME24	AIME25	GPQA
0.0	79.3/88.7	12.1/27.0	11.1/34.8	37.9/81.4
0.5	**79.8/93.8**	**19.0/43.2**	16.1/**41.9**	**38.8/89.1**
1.0	77.4/93.5	17.1/35.7	**16.8**/36.1	35.4/81.7
1.5	59.6/89.0	5.8/24.0	7.0/31.2	33.5/80.2
2.0	36.4/84.4	4.4/16.4	4.3/23.0	33.0/84.3

Table 17: Sensitivity to mixing coefficient α. Qwen3-4B-Base trained on MATH-500. Each cell shows `pass@1`/`pass@16`.

α	MATH	AIME24	AIME25	AMC	GPQA
0.00	78.4/92.4	17.1/39.8	14.4/38.0	48.7/79.8	37.8/87.3
0.25	79.4/93.4	18.5/42.1	15.7/40.9	49.9/81.7	38.5/88.6
0.50	79.8/93.8	19.0/43.2	16.1/41.9	50.3/82.2	38.8/89.1
0.75	79.5/93.6	18.6/42.6	15.9/41.5	50.1/82.0	38.6/88.9
1.00	79.1/93.0	18.2/41.4	15.3/40.2	49.4/81.2	38.3/88.4

C Additional Rationale Supporting the Reward Design

C.1 Analysis of Reward Formulation on "Almost Correct" Solutions

A key nuance in our reward formulation is how it handles solutions that are "almost correct"—for instance, a minority solution that is semantically similar to a correct majority path but fails on a minor step. This section provides a detailed analysis of how our two-part novelty score, $u_i = 1 - (\alpha\bar{s}_i + (1-\alpha)m_i)$, is specifically designed to handle this nuance.

Our primary goal is to apply strong negative pressure against high-frequency, common failure modes, not to indiscriminately punish all incorrect explorations. In the scenario where a minority solution is semantically similar to the majority (resulting in a high global similarity m_i), the intra-group mean similarity $\bar{s}_i$ (the solution's average similarity to other minority solutions) becomes the critical differentiator. We analyze two distinct cases:

- **Case 1: The error is a rare, exploratory mistake.** If the "almost correct" error is a rare, occasional mistake, the reasoning path will be semantically dissimilar from the other failure modes in the minority group. This results in a low $\bar{s}_i$. According to our novelty formula, this low $\bar{s}_i$ counteracts the high m_i, leading to a higher overall novelty score u_i. This effectively mitigates the penalty, ensuring the model is not strongly discouraged from valid exploration.
- **Case 2: The error is a common failure mode.** If the error represents a high-frequency failure mode (e.g., a consistent arithmetic error), the reasoning path will be semantically similar to many other solutions in the minority group. This results in a high $\bar{s}_i$. In this case, both the $\bar{s}_i$ and m_i terms are high, leading to a very low u_i. This results in a maximum penalty, preventing the model from collapsing into a "consistent but wrong" state.

This design ensures that our reward mechanism is robust. It relies on the group-level context to selectively protect valuable explorations while aggressively pruning systematic errors. This is particularly crucial in high-uncertainty scenarios where the novelty signal must accurately guide exploration, which is the exact behavior we aim to encourage.

D Theoretical Justification

In this section, we formally justify the different behaviors of the optimal policy under correctness-only reinforcement learning and under our variation-augmented objective. We show that, for a suitably small variation weight and under a simple dissimilarity-gap structure, both objectives have globally optimal policies that concentrate on the set of correct reasoning traces, but our variation-augmented

Table 18: Sensitivity to sampling temperature T for the trained EVOL-RL model (Qwen3-4B-Base, MATH-TRAIN). Each cell shows `pass@1`/`pass@16`.

T	MATH	AIME24	AIME25	AMC	GPQA
0.0	82.0/82.0	23.3/23.3	16.7/16.7	54.2/54.2	36.0/36.0
0.2	80.7/92.9	20.2/42.2	16.6/39.9	52.7/80.6	39.0/84.7
0.4	80.3/93.1	20.2/40.5	16.4/41.4	52.3/79.5	38.7/87.2
0.6	80.3/93.4	19.3/42.3	18.3/42.3	51.5/80.1	38.2/89.0
0.8	79.9/93.6	19.6/43.1	18.5/40.2	52.1/81.3	37.6/91.2
1.0	79.3/93.8	18.9/42.6	16.0/43.4	51.5/83.4	37.5/91.3

objective further selects solutions with strictly more diverse coverage over correct trajectories and, under additional symmetry condition, maximal policy entropy on the correct set.

D.1 SETUP

For a given reasoning question q, there is a finite set $\mathcal{Y}$ of complete chain-of-thought (CoT) trajectories for q. Each trajectory receives a binary task reward

$$r(y) \in \{0, 1\}, \qquad y \in \mathcal{Y}.$$

We denote the correct set (answer-correct set) by

$$G := \{y \in \mathcal{Y} : r(y) = 1\},$$

and assume $G \neq \emptyset$. A policy π maps question to chain of thought reasoning. For simplicity, we will omit the condition of question in the following analysis.

Objectives. The correctness-only objective is

$$J_0(\pi) := \mathbb{E}_\pi[r(Y)] = \sum_{y \in \mathcal{Y}} \pi(y)\, r(y).$$

To encourage diverse correct solutions, we introduce a symmetric, nonnegative CoT-level dissimilarity

$$d : \mathcal{Y} \times \mathcal{Y} \to [0, \infty), \quad d(y, y') = d(y', y), \quad d(y, y) = 0,$$

and assume that d is bounded:

$$D_{\max} := \max_{y, y' \in \mathcal{Y}} d(y, y') < \infty. \tag{2}$$

Given a group size $K \geq 2$ and variation weight $\lambda > 0$, we form i.i.d. samples $Y_1, \ldots, Y_K \sim \pi$ and define the group reward

$$R_{\text{group}}(Y_{1:K}) := \frac{1}{K} \sum_{k=1}^{K} r(Y_k) + \lambda \cdot \frac{1}{K(K-1)} \sum_{k \neq \ell} d(Y_k, Y_\ell).$$

The correctness-plus-variation objective is the expected group reward

$$J_\lambda(\pi) := \mathbb{E}_{Y_{1:K} \sim \pi}[R_{\text{group}}(Y_{1:K})].$$

The following standard computation shows that J_λ is a quadratic functional of π.

Lemma 1 (Expected group reward as a quadratic in π)**.** *For any $\pi \in \Delta(\mathcal{Y})$,*

$$J_\lambda(\pi) = \sum_y \pi(y)\, r(y) + \lambda \sum_{y, y'} \pi(y)\pi(y')\, d(y, y'). \tag{3}$$

Proof. By linearity of expectation and the i.i.d. assumption on $Y_{1:K}$, the first term yields

$$\mathbb{E}\Big[\frac{1}{K} \sum_{k=1}^{K} r(Y_k)\Big] = \mathbb{E}[r(Y_1)] = \sum_y \pi(y) r(y).$$

For the second term, each ordered pair (k, ℓ) with $k \neq \ell$ has the same distribution, so

$$\mathbb{E}\Big[\frac{1}{K(K-1)} \sum_{k \neq \ell} d(Y_k, Y_\ell)\Big] = \mathbb{E}[d(Y_1, Y_2)] = \sum_{y, y'} \pi(y)\pi(y') d(y, y'),$$

which gives (3). □

D.2 Step 1: Global maximizers concentrate on the correct set

We first show that for a suitable range of λ, any global maximizer of J_0 or J_λ puts zero probability on incorrect trajectories.

Lemma 2 (Global maximizers of J_0 are supported on G)**.** *Assume $G \neq \emptyset$. Then any global maximizer $\pi^\star$ of J_0 satisfies*

$$\mathrm{supp}(\pi^\star) \subseteq G.$$

Moreover, the set of global maximizers of J_0 is exactly $\Delta(G)$.

Proof. For any $\pi \in \Delta(\mathcal{Y})$,

$$J_0(\pi) = \sum_y \pi(y) r(y) = \sum_{y \in G} \pi(y) = \pi(G) \leq 1,$$

with equality iff $\pi(G) = 1$, i.e., $\mathrm{supp}(\pi) \subseteq G$. Because $G \neq \emptyset$, there exists a policy with $\pi(G) = 1$ and $J_0(\pi) = 1$, so the maximal value of J_0 is 1, achieved exactly by policies supported on G. □

For the variation-augmented objective, the task reward gap between correct and incorrect trajectories is 1, while the variation term is bounded by $D_{\max}$ from (2). For sufficiently small λ, the former always dominates the latter.

Lemma 3 (For small λ, global maximizers of J_λ are supported on G)**.** *Suppose $G \neq \emptyset$ and $D_{\max}$ is defined as in (2). If*

$$0 < \lambda < \frac{1}{2D_{\max}}, \tag{4}$$

(with the convention that the condition is vacuous if $D_{\max} = 0$), then any global maximizer $\pi^\star$ of J_λ satisfies

$$\mathrm{supp}(\pi^\star) \subseteq G.$$

Proof. Using (3), we can write

$$J_\lambda(\pi) = \sum_y \pi(y) r(y) + \lambda \sum_{y,y'} \pi(y)\pi(y') d(y, y').$$

For $y \in \mathcal{Y}$, the partial derivative of J_λ w.r.t. $\pi(y)$ is

$$g_y(\pi) := \frac{\partial J_\lambda(\pi)}{\partial \pi(y)} = r(y) + 2\lambda \sum_{y'} \pi(y') d(y, y'),$$

where we used the symmetry $d(y, y') = d(y', y)$.

Assume for contradiction that $\pi^\star$ is a global maximizer of J_λ and there exists an incorrect trajectory $y_0 \notin G$ with $\pi^\star(y_0) > 0$. Because $G \neq \emptyset$, pick any $y_1 \in G$ (so $r(y_1) = 1$). Consider the feasible direction in the simplex

$$v := e_{y_1} - e_{y_0},$$

which corresponds to infinitesimally moving probability mass from y_0 to y_1. For sufficiently small $\varepsilon > 0$, the perturbed policy $\pi^\star + \varepsilon v$ remains in $\Delta(\mathcal{Y})$.

The directional derivative of J_λ at $\pi^\star$ along v is

$$\frac{d}{d\varepsilon} J_\lambda(\pi^\star + \varepsilon v)\Big|_{\varepsilon=0} = \langle \nabla J_\lambda(\pi^\star), v \rangle = g_{y_1}(\pi^\star) - g_{y_0}(\pi^\star).$$

By definition of g_y,

$$\begin{aligned} g_{y_1}(\pi^\star) - g_{y_0}(\pi^\star) &= \big(r(y_1) - r(y_0)\big) + 2\lambda \sum_{y'} \pi^\star(y') \big(d(y_1, y') - d(y_0, y')\big) \\ &= 1 + 2\lambda \sum_{y'} \pi^\star(y') \big(d(y_1, y') - d(y_0, y')\big), \end{aligned}$$

since $r(y_1) = 1$ and $r(y_0) = 0$. By the definition of $D_{\max}$,

$$\big|d(y_1, y') - d(y_0, y')\big| \le D_{\max} \quad \text{for all } y',$$

and since $\sum_{y'} \pi^\star(y') = 1$, we obtain

$$\left|\sum_{y'} \pi^\star(y')\big(d(y_1, y') - d(y_0, y')\big)\right| \le D_{\max}.$$

Hence

$$g_{y_1}(\pi^\star) - g_{y_0}(\pi^\star) \ge 1 - 2\lambda D_{\max}.$$

If $0 < \lambda < 1/(2D_{\max})$, then $1 - 2\lambda D_{\max} > 0$, so

$$\langle \nabla J_\lambda(\pi^\star), v\rangle > 0.$$

Thus moving a small amount of mass from y_0 to y_1 strictly increases J_λ, contradicting the assumption that $\pi^\star$ is a (global, hence local) maximizer. Therefore no maximizer can assign positive probability to any incorrect trajectory, and $\mathrm{supp}(\pi^\star) \subseteq G$. □

Combining Lemma 2 and Lemma 3, we obtain that assume $G \neq \emptyset$, $D_{\max} < \infty$, and (4). Then any global maximizer of J_0 or J_λ is supported on G. Thus we may restrict attention to

$$\Delta(G) := \{\pi \in \Delta(\mathcal{Y}) : \mathrm{supp}(\pi) \subseteq G\}$$

when comparing optimal policies.

D.3 Step 2: Mode structure inside G and a piecewise-homogeneous dissimilarity gap

We now formalize the assumption that the correct set G contains multiple qualitatively distinct solution modes, with large dissimilarity gaps between modes and relatively homogeneous dissimilarity within each mode.

Assumption 1 (Mode partition of correct CoTs)**.** *The correct set G is partitioned into $M \ge 2$ disjoint subsets*

$$G = G_1 \cup \cdots \cup G_M, \qquad G_m \cap G_{m'} = \emptyset \ (m \neq m'),$$

where each G_m corresponds to a distinct reasoning mode (solution pattern). Let $N_m := |G_m|$ and $N := |G| = \sum_{m=1}^{M} N_m$.

For any $\pi \in \Delta(G)$, we define the total probability mass on each mode:

$$w_m(\pi) := \sum_{y \in G_m} \pi(y), \qquad m = 1, \ldots, M.$$

Then $w(\pi) = (w_1(\pi), \ldots, w_M(\pi))$ lies in the simplex

$$\Delta_M := \{w \in \mathbb{R}_+^M : \sum_{m=1}^{M} w_m = 1\}.$$

We impose a piecewise-homogeneous dissimilarity-gap assumption on d within G.

Assumption 2 (Piecewise-homogeneous dissimilarity on the correct set)**.** *There exist constants $D_{\mathrm{in}} \ge 0$ and $D_{\mathrm{out}} > D_{\mathrm{in}}$ such that for all $y, y' \in G$,*

$$d(y, y') = \begin{cases} 0, & \text{if } y = y', \\ D_{\mathrm{in}}, & \text{if } y \neq y',\ y, y' \in G_m \text{ for some } m \text{ (same mode)}, \\ D_{\mathrm{out}}, & \text{if } y \in G_m,\ y' \in G_{m'}\ (m \neq m') \text{ (different modes)}. \end{cases}$$

Assumption 2 idealizes the intuitive condition that trajectories within the same mode are relatively similar (with dissimilarity D_{in}), while trajectories from different modes are more dissimilar (with $D_{\mathrm{out}} > D_{\mathrm{in}}$).

Under these assumptions, the variation term in J_λ over $\Delta(G)$ admits a convenient decomposition into a mode-level and an intra-mode component.

Lemma 4 (Decomposition of the dissimilarity term on G)**.** *Under Assumptions 1 and 2, for any* $\pi \in \Delta(G)$*,*

$$\sum_{y,y' \in G} \pi(y)\pi(y')d(y,y') = D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_{m=1}^{M} w_m(\pi)^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2.$$

Consequently, on $\Delta(G)$*,*

$$J_\lambda(\pi) = 1 + \lambda \Big[D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_{m=1}^{M} w_m(\pi)^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2 \Big]. \tag{5}$$

Proof. Because $\text{supp}(\pi) \subseteq G$, we have $r(y) = 1$ on G and $r(y) = 0$ otherwise, hence $\sum_y \pi(y)r(y) = \sum_{y \in G} \pi(y) = 1$, giving the first term in (5).

For the dissimilarity term, we decompose by modes:

$$\begin{aligned} \sum_{y,y' \in G} \pi(y)\pi(y')d(y,y') = &\sum_{m=1}^{M} \sum_{y,y' \in G_m} \pi(y)\pi(y')d(y,y') \\ &+ \sum_{m \neq m'} \sum_{y \in G_m,\, y' \in G_{m'}} \pi(y)\pi(y')d(y,y'). \end{aligned}$$

Within a fixed mode G_m, we have $d(y,y) = 0$ and $d(y,y') = D_{\text{in}}$ for $y \neq y'$, so

$$\begin{aligned} \sum_{y,y' \in G_m} \pi(y)\pi(y')d(y,y') &= D_{\text{in}} \sum_{\substack{y,y' \in G_m \\ y \neq y'}} \pi(y)\pi(y') \\ &= D_{\text{in}} \Big[\Big(\sum_{y \in G_m} \pi(y) \Big)^2 - \sum_{y \in G_m} \pi(y)^2 \Big] \\ &= D_{\text{in}} \Big[w_m(\pi)^2 - \sum_{y \in G_m} \pi(y)^2 \Big]. \end{aligned}$$

Summing over m yields

$$\sum_{m=1}^{M} \sum_{y,y' \in G_m} \pi(y)\pi(y')d(y,y') = D_{\text{in}} \sum_{m=1}^{M} w_m(\pi)^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2.$$

Across different modes, $d(y,y') = D_{\text{out}}$ whenever $y \in G_m$ and $y' \in G_{m'}$ with $m \neq m'$, hence

$$\sum_{m \neq m'} \sum_{y \in G_m,\, y' \in G_{m'}} \pi(y)\pi(y')d(y,y') = D_{\text{out}} \sum_{m \neq m'} w_m(\pi) w_{m'}(\pi).$$

Since $\sum_m w_m(\pi) = 1$,

$$\sum_{m \neq m'} w_m w_{m'} = \Big(\sum_m w_m \Big)^2 - \sum_m w_m^2 = 1 - \sum_m w_m^2.$$

Combining the within-mode and cross-mode contributions gives

$$\begin{aligned} \sum_{y,y' \in G} \pi(y)\pi(y')d(y,y') &= D_{\text{in}} \sum_m w_m^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2 + D_{\text{out}} \Big(1 - \sum_m w_m^2 \Big) \\ &= D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_m w_m^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2. \end{aligned}$$

Substituting this into Lemma 1 with $\sum_y \pi(y)r(y) = 1$ yields (5). □

Thus, over the feasible region $\Delta(G)$, maximizing $J_\lambda(\pi)$ is equivalent to maximizing

$$D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_{m=1}^{M} w_m(\pi)^2 - D_{\text{in}} \sum_{y \in G} \pi(y)^2,$$

or, equivalently (since $\lambda > 0$ and constants do not affect argmax), to minimizing

$$F(\pi) := (D_{\text{out}} - D_{\text{in}}) \sum_{m=1}^{M} w_m(\pi)^2 + D_{\text{in}} \sum_{y \in G} \pi(y)^2, \tag{6}$$

where $D_{\text{out}} - D_{\text{in}} > 0$ and $D_{\text{in}} \geq 0$.

D.4 Step 3: Coverage structure and entropy of optimal policies on G

Having reduced both objectives to $\Delta(G)$ (Lemma 2 and Lemma 3), we now compare their optimal solutions. First, Lemma 2 immediately implies:

Corollary 1 (Flat optimal set for correctness-only objective)**.** *Under $G \neq \emptyset$, any $\pi \in \Delta(G)$ satisfies $J_0(\pi) = 1$ and is a global maximizer of J_0. Hence J_0 is indifferent to how probability mass is distributed within G.*

In contrast, J_λ has a nontrivial preference both at the mode level and within each mode, captured by the quadratic form $F(\pi)$ in (6).

We first show that, for any fixed mode-mass vector w, J_λ is maximized by making the policy uniform within each mode.

Lemma 5 (Within each mode, optimal policies are uniform)**.** *Fix $w \in \Delta_M$ and consider the set*

$$\mathcal{P}(w) := \{\pi \in \Delta(G) : w_m(\pi) = w_m \text{ for all } m\}.$$

Under Assumptions 1 and 2, among all $\pi \in \mathcal{P}(w)$, $J_\lambda(\pi)$ is maximized (equivalently, $F(\pi)$ is minimized) by policies that are uniform within each mode:

$$\pi(y) = \frac{w_m}{N_m} \quad \text{for all } y \in G_m.$$

If $D_{\text{in}} > 0$, this choice is unique in $\mathcal{P}(w)$; if $D_{\text{in}} = 0$, J_λ is independent of the within-mode distribution.

Proof. For fixed w, the term $\sum_m w_m(\pi)^2$ in (6) is constant over $\mathcal{P}(w)$. Thus minimizing $F(\pi)$ over $\mathcal{P}(w)$ is equivalent to minimizing $\sum_{y \in G} \pi(y)^2$ over $\mathcal{P}(w)$.

Within each mode G_m, this is the classical problem of minimizing a sum of squares subject to a linear constraint:

$$\min\Big\{ \sum_{y \in G_m} \pi(y)^2 : \sum_{y \in G_m} \pi(y) = w_m,\ \pi(y) \geq 0 \Big\},$$

whose unique solution (when $w_m > 0$) is the uniform allocation $\pi(y) = w_m / N_m$ for all $y \in G_m$. If $D_{\text{in}} > 0$, this sum of squares enters $F(\pi)$ with a strictly positive coefficient, so any deviation from uniformity strictly increases $F(\pi)$. If $D_{\text{in}} = 0$, the intra-mode term vanishes from $F(\pi)$, and $F(\pi)$ (hence J_λ) depends only on w and not on the within-mode distribution. □

By Lemma 5, any global maximizer of J_λ on $\Delta(G)$ must be uniform within each mode. We may therefore restrict attention to policies of the form

$$\pi(y) = \frac{w_m}{N_m} \quad \text{if } y \in G_m,$$

parameterized solely by $w \in \Delta_M$. Substituting this structure into (5) yields a purely mode-level objective.

Indeed, for such a π,

$$\sum_{y \in G} \pi(y)^2 = \sum_{m=1}^{M} \sum_{y \in G_m} \left(\tfrac{w_m}{N_m}\right)^2 = \sum_{m=1}^{M} \frac{w_m^2}{N_m}.$$

Plugging this into Lemma 4 gives

$$J_\lambda(\pi) = 1 + \lambda\Big[D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_{m=1}^{M} w_m^2 - D_{\text{in}} \sum_{m=1}^{M} \frac{w_m^2}{N_m}\Big]$$
$$= 1 + \lambda\Big[D_{\text{out}} - \sum_{m=1}^{M} a_m w_m^2\Big], \tag{7}$$

where we define

$$a_m := (D_{\text{out}} - D_{\text{in}}) + \frac{D_{\text{in}}}{N_m} > 0.$$

Thus, among mode-wise uniform policies, maximizing J_λ is equivalent to minimizing

$$\sum_{m=1}^{M} a_m w_m^2 \quad \text{subject to } w \in \Delta_M.$$

Theorem 2 (Variation promotes broader mode coverage on G)**.** *Suppose $G \neq \emptyset$, Assumptions 1 and 2 hold with $M \geq 2$, and λ satisfies (4). Let $\pi^{(\lambda)}$ be any global maximizer of J_λ over $\Delta(G)$, and let $w^{(\lambda)} := w(\pi^{(\lambda)})$ be its mode-mass vector. Then:*

1. *$\pi^{(\lambda)}$ is uniform within each mode:*

$$\pi^{(\lambda)}(y) = \frac{w_m^{(\lambda)}}{N_m} \quad \text{for all } y \in G_m,\ m = 1, \dots, M.$$

2. *$w^{(\lambda)}$ is the unique minimizer of $\sum_m a_m w_m^2$ over Δ_M:*

$$w_m^{(\lambda)} = \frac{a_m^{-1}}{\sum_{j=1}^{M} a_j^{-1}}, \quad m = 1, \dots, M,$$

where $a_m > 0$ is defined above. In particular, every mode receives strictly positive probability:

$$w_m^{(\lambda)} > 0 \quad \text{for all } m = 1, \dots, M.$$

3. *If, in addition, $D_{\text{in}} > 0$ and all modes have equal size $N_m \equiv N/M$, then all a_m coincide, $w^{(\lambda)}$ is uniform on modes, and $\pi^{(\lambda)}$ is the uniform distribution over all correct trajectories:*

$$\pi^{(\lambda)}(y) = \frac{1}{|G|} \quad \text{for all } y \in G.$$

In this symmetric case, $\pi^{(\lambda)}$ maximizes the full policy entropy

$$H(\pi) := -\sum_{y \in G} \pi(y) \log \pi(y),$$

achieving $H(\pi^{(\lambda)}) = \log |G|$, the largest possible entropy on $\Delta(G)$.

In contrast, any $\pi^{(0)} \in \Delta(G)$ is optimal for J_0, including degenerate solutions that concentrate on a single mode (with mode-level entropy 0 and very low trajectory-level entropy). Thus J_λ provably selects *high-coverage policies within G, using all modes and being uniform within each mode, and, under mild symmetry, the maximum-entropy fully uniform policy, while J_0 does not enforce any coverage.*

Proof. (i) By Lemma 3, any global maximizer $\pi^{(\lambda)}$ lies in $\Delta(G)$. Lemma 5 then implies that any maximizer must be uniform within each mode, so $\pi^{(\lambda)}$ has the stated form.

(ii) For such mode-wise uniform policies, (7) shows that maximizing J_λ is equivalent to minimizing $\sum_m a_m w_m^2$ over $w \in \Delta_M$. The function $w \mapsto \sum_m a_m w_m^2$ is strictly convex on Δ_M because each

$a_m > 0$, and the constraint set Δ_M is convex and compact. Using Lagrange multipliers for the equality constraint $\sum_m w_m = 1$, the unique minimizer satisfies

$$2a_m w_m^{(\lambda)} + \mu = 0 \quad \text{for all } m,$$

for some scalar μ, which yields

$$w_m^{(\lambda)} = \frac{-\mu}{2a_m} = \frac{a_m^{-1}}{\sum_{j=1}^{M} a_j^{-1}}, \quad m = 1, \ldots, M.$$

All $a_m > 0$, so all $w_m^{(\lambda)} > 0$.

(iii) If $D_{\text{in}} > 0$ and $N_m \equiv N/M$ for all m, then

$$a_m = (D_{\text{out}} - D_{\text{in}}) + \frac{D_{\text{in}}}{N_m} = (D_{\text{out}} - D_{\text{in}}) + \frac{D_{\text{in}}}{N/M}$$

is the same for all m. Hence $w_m^{(\lambda)} = 1/M$ for all m, and

$$\pi^{(\lambda)}(y) = \frac{w_m^{(\lambda)}}{N_m} = \frac{1/M}{N/M} = \frac{1}{N} = \frac{1}{|G|}$$

for all $y \in G$, i.e., $\pi^{(\lambda)}$ is the uniform distribution over G.

Finally, the Shannon entropy $H(\pi)$ over G is maximized on $\Delta(G)$ exactly by the uniform distribution, with value $\log |G|$. Thus $H(\pi^{(\lambda)}) = \log |G|$, the largest possible entropy on $\Delta(G)$. □

Summary. Under the binary-reward setting and a piecewise-homogeneous dissimilarity-gap structure on the correct set, we have shown that for any sufficiently small variation weight λ:

(a) (Lemma 2 and Lemma 3) Any global maximizer of J_0 or J_λ places all its probability mass on the correct set G; i.e., both objectives "converge to the answer-correct set" in a static sense.

(b) (Lemma 1) The correctness-only objective J_0 is completely flat on $\Delta(G)$: every distribution over G is globally optimal, including degenerate policies that collapse onto a single mode and assign zero probability to many correct trajectories.

(c) (Theorem 2) In contrast, the variation-augmented objective J_λ has global maximizers that must put positive mass on *all* modes and are uniform within each mode, and in the symmetric case where all modes have equal size, it selects the uniform distribution over all correct trajectories, which maximizes the full policy entropy.

Consequently, if a training procedure converges to global maximizers of J_0 and J_λ, then in the binary setting with dissimilarity gaps: both trainings converge to fully correct policies, but the variation-augmented training *provably selects* strictly more dispersed solutions on the correct set G, using all modes and, under mild symmetry, the maximum-entropy policy—while correctness-only training does not favor coverage and may converge to highly collapsed solutions.

E Use of Large Language Models in Preparation

We acknowledge the use of Large Language Models (LLMs) as assistants in the preparation of this manuscript. Their role included refining phrasing and improving the clarity of the text, as well as assisting with programming tasks such as code generation and debugging for our experiments. The authors critically reviewed, edited, and verified all LLM-generated content for accuracy and appropriateness, and take full responsibility for the final content of this paper.

F Limitations

Our study has several scope limitations. First, Evol-RL preserves alternative solutions that may enable later correction, but it does not guarantee that a persistent incorrect majority will be overturned.

Second, our training experiments primarily focus on mathematical reasoning; while we observe positive transfer to several held-out reasoning benchmarks, broader domain generalization remains to be studied. Finally, embedding-based novelty provides a practical measure of variation rather than a complete characterization of reasoning diversity. We therefore complement it with analyses of corrective minority solutions and distinct successful reasoning strategies.