

Learn from Zero: Policy Optimization under Vanishing Advantage

Chang Liu^{1*}Peng Xiao¹Jiandong Su¹Wayne Zhang^{1,2†}

liuchang0712@buaa.edu.cn

xiaopeng1@sensetime.com

jiandong.laurence.su@gmail.com

wayne.zhang@sensetime.com

¹SenseTime Research²Shanghai Key Laboratory of Quantum City Spatial Intelligence Innovation

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a vital paradigm for improving the reasoning capabilities of Large Language Models (LLMs). However, methods such as Group Relative Policy Optimization (GRPO) provide no effective learning signal when all responses sampled for a prompt receive the same reward. This zero-variance setting causes the group advantage to vanish and discards potentially useful training signals, thereby limiting both exploration and reasoning performance. To address this issue, we propose **LZPO**, an algorithm that recovers learning signals from zero-variance prompts. For *all_correct* groups, LZPO assigns semantic-entropy-guided advantages to encourage exploration among successful reasoning paths. For *all_wrong* groups, it generates a reflection response and ranks the original erroneous responses according to their semantic alignment with this response. The reflection response is used only for relative ranking; it is neither treated as verified ground truth nor assigned directly as a positive training target. Together, these mechanisms recover learning signals that GRPO otherwise discards. Experiments on six mathematical reasoning benchmarks show that LZPO achieves gains of up to 3.3 points over GRPO and outperforms other baselines, with particularly pronounced improvements on challenging reasoning tasks.

1 Introduction

Reinforcement Learning (RL) (Schulman et al., 2017; Shao et al., 2024; Yue et al., 2025) serves as a critical framework for post-training Large Language Models (LLMs). Among its paradigms, Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI et al., 2025; Su et al., 2025) leverages rule-based evaluation to enhance reasoning without learned reward models, while Group Relative Policy Optimization (GRPO) (Shao et al., 2024) further improves efficiency by eliminating the value network and achieves better performance on complex reasoning tasks.

However, GRPO-based methods face a significant bottleneck with zero-variance prompts, where all sampled responses yield identical rewards. As shown in Figure 1 Right, these prompts comprise 40%–60% of training data. In such scenarios, the group advantage becomes zero, causing gradients to vanish and severely degrading sample efficiency. Existing solutions such as DAPO (Yu et al., 2025; Zheng et al., 2025; Liao et al., 2025) attempt to filter these prompts but incur substantial computational overhead, with rollouts taking up nearly 50% of the training time (Le et al., 2025). Moreover, discarding this data exacerbates signal sparsity in later training stages. This limits exploration (Wang et al., 2025; Xie et al., 2025; Deng et al., 2025) and increases the risk of entropy collapse (Cui et al., 2025), ultimately constraining model performance.

Contrary to prior assumptions (Yu et al., 2025; Zheng et al., 2025; Liao et al., 2025), we argue that zero-variance prompts contain valuable latent training signals, as shown in Figure 1 Left. Effectively exploiting these signals is crucial for preserving learned capabilities and guiding exploration toward diverse reasoning paths, even in the absence of explicit reward

¹<https://github.com/love-cfall/LZPO>

*Work done during internship at SenseTime.

†Corresponding author: Wayne Zhang.

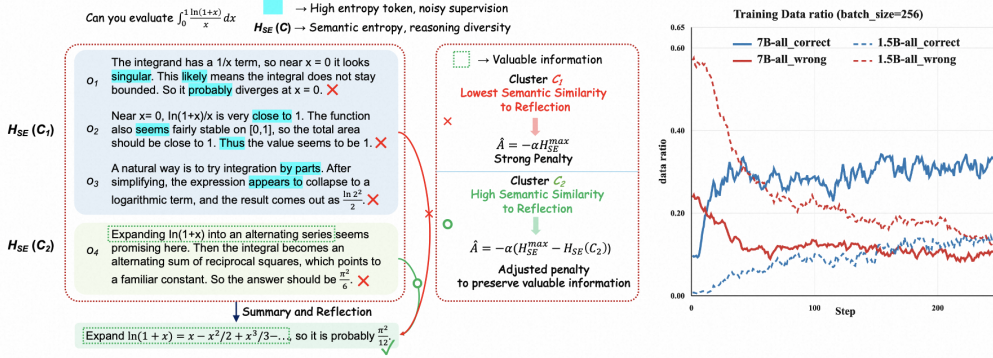


Figure 1: Semantic diversity in zero-variance prompts. **Left:** *all_wrong* responses exhibit varying semantic entropy, and reflection provides useful guidance for difficult problems. **Right:** evolution of zero-variance prompt composition (*all_correct* vs. *all_wrong*) during training.

differences. Recent work (Le et al., 2025) extracts learning signals from zero-variance prompts with token-level entropy. However, we find that semantic entropy is less noisy than token-level entropy because it captures diversity in underlying reasoning meanings rather than local token fluctuations (Huang et al., 2025; Chen et al., 2025a).

To this end, we propose **LZPO** (Learn from Zero Policy Optimization), a novel algorithm that reactivates training signals through two complementary mechanisms: (1) a semantic-entropy-based advantage that encourages exploration by assigning higher relative advantages to *all_correct* responses with greater intra-group semantic diversity; (2) a reflection-guided mechanism that, for *all_wrong* responses, summarizes errors, generates a reflection response, and uses it as a weak directional signal to distinguish informative mistakes from uninformative failures.

Our main contributions are as follows. We observe that *zero-variance prompts* contain useful latent training signals and propose to exploit them through semantic entropy and reflection-guided mechanisms to mitigate vanishing gradients and improve sample efficiency in RL training. We propose **LZPO**, a policy optimization algorithm that reactivates these signals through two complementary mechanisms. Experiments on six mathematical reasoning benchmarks across multiple model scales show that LZPO consistently outperforms GRPO baselines while incurring conditional computation from reflection generation on *all_wrong* prompt groups.

2 Related work

2.1 Reinforcement learning for LLMs

In recent years, reinforcement learning has become a core paradigm for enhancing the reasoning capabilities of large language models. DeepSeek-AI (DeepSeek-AI et al., 2025) and OpenAI (OpenAI et al., 2024) show that reward-based optimization on verifiable tasks can substantially improve the stability and generalization of model reasoning. By constructing group-relative advantages, GRPO (Shao et al., 2024) removes the reliance on complex critic models and achieves better performance gains on reasoning tasks. To address the limitations of GRPO, several follow-up methods have been proposed. Approaches (Yu et al., 2025; Wang et al., 2025; Su et al., 2025; Cui et al., 2025) refine the clipping strategy to better optimize high-entropy tokens, mitigating entropy collapse during training. In parallel, other approaches (Chen et al., 2025b; Wu et al., 2025a; Zhang et al., 2025) redesign the advantage function to explicitly encourage exploration, further improving reasoning performance. However, existing methods primarily focus on non-zero-variance samples and largely overlook the informative signals embedded in zero-variance data. In contrast, our approach reactivates latent learning signals from zero-variance prompts, enabling more efficient utilization of both training data and computational resources.

2.2 Data sampling and advantage shaping strategies

Addressing the challenge of zero-variance prompts in RL, **Data sampling strategies** primarily improve training efficiency by filtering or downsampling such samples. DAPO (Yu et al., 2025) introduces a dynamic sampler to ensure that each batch contains only non-zero-variance samples, while other methods (Xiong et al., 2025; Liao et al., 2025; Cao et al., 2025; Zheng et al., 2025) reduce the sampling frequency of zero-variance prompts and adopt curriculum learning to filter data during training. **Advantage shaping strategies** instead recover learning signals from zero-variance samples. ZVP (Le et al., 2025) assigns token-level advantages using token-level characteristics, GCPO (Wu & Liu, 2025) activates advantages through teacher-trajectory distillation and intra-group replacement, and SEED-GRPO (Chen et al., 2025a) uses prompt-level semantic entropy to scale updates conservatively for uncertain prompts. In contrast, LZPO constructs response-level advantages from semantic clusters: cluster-level entropy guides *all_correct* groups, whereas reflection-guided alignment ranks erroneous responses in *all_wrong* groups, directly targeting vanishing group advantages.

3 Preliminary

3.1 Group relative policy optimization

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is an extension of PPO (Schulman et al., 2017) that stabilizes policy updates via group-wise normalization of advantage estimates, without requiring an explicit value function. Given a question-answer pair $(q, a) \sim \mathcal{D}$ sampled from the training dataset, the old policy π_{old} generates G independent responses $\{o_i\}_{i=1}^G$ for the same query q . The GRPO objective is defined as

$$\mathcal{J}_{\text{GRPO}}(\theta \mid q, a, \{o_i\}_{i=1}^G) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t}) \quad (1)$$

where $r_{i,t}(\theta)$ denotes the importance sampling ratio:

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} \mid q, o_{i,<t})}{\pi_{\text{old}}(o_{i,t} \mid q, o_{i,<t})}. \quad (2)$$

The token-level advantage $\hat{A}_{i,t}$ is computed using group-normalized rewards:

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \quad (3)$$

Here, r_i denotes the scalar reward associated with the complete response o_i . By normalizing rewards across responses to the same prompt, GRPO captures relative performance within each group, guiding the policy toward comparatively better solutions.

3.2 Semantic entropy

Semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) measures uncertainty at the *semantic level* by quantifying diversity in meaning rather than surface-level token variation. Given a query q , we sample a group of G responses from the previous policy:

$$\{o_1, o_2, \dots, o_G\} \sim \pi_{\text{old}}(\cdot \mid q). \quad (4)$$

Following prior work, we estimate semantic entropy from sampled responses using a Monte Carlo approximation (Kossen et al., 2024). Specifically, each response o_j is represented by a sequence-level embedding obtained by mean pooling the last-layer hidden states:

$$h_{o_j} = \frac{1}{|o_j|} \sum_{t=1}^{|o_j|} h_{j,t}. \quad (5)$$

In our implementation, these rollout embeddings are derived from vLLM hidden states and clustered with K-means using a fixed $K = 2$; sensitivity to K is studied in Section 5.4.

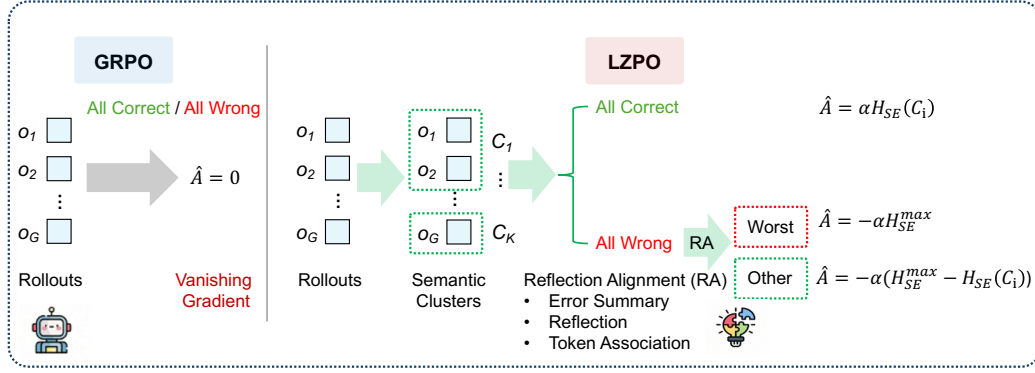


Figure 2: Overview of LZPO advantage calculation with semantic entropy and reflection.

We then apply K-means to $\{h_{o_j}\}_{j=1}^G$, yielding K semantic clusters $\mathcal{C} = \{C_1, \dots, C_K\}$, where each cluster contains semantically similar responses. We estimate the probability mass of cluster C_i from the sampled group as

$$P(C_i | q) = \frac{1}{G} \sum_{j=1}^G \mathbf{1}(o_j \in C_i), \quad i = 1, \dots, K. \quad (6)$$

The entropy contribution of cluster C_i is then $H_{SE}(C_i) = -P(C_i | q) \log P(C_i | q)$, and the overall semantic entropy is $H_{SE}(q) = \sum_{i=1}^K H_{SE}(C_i) = -\sum_{i=1}^K P(C_i | q) \log P(C_i | q)$. In Section 4.2, we further use this cluster-level entropy contribution to guide response-level advantage assignment.

4 Method

4.1 Semantic entropy guided advantage design for *all_correct* prompt

When all responses in a group are either correct or wrong, GRPO yields zero advantages and thus vanishing policy gradients in Eq. 7, making the prompt ineffective for policy optimization.

$$\nabla_{\theta} \mathcal{J}_{\text{GRPO}} \propto \mathbb{E}[\hat{A}_{i,t} \nabla_{\theta} \log \pi_{\theta}(o_{i,t} | q, o_{i,<t})]. \quad (7)$$

We denote prompts whose group responses are all correct or all wrong at each training step as *all_correct* and *all_wrong*. Empirically, these zero-variance prompts account for nearly 45% of training, as shown in Figure 1 Right, indicating that existing training paradigms waste a substantial amount of data. Our key insight is that zero-variance prompts are not uninformative, but require finer-grained signals to be exploited effectively. Semantic entropy reveals latent structural differences and the exploration space beneath superficially uniform group responses, providing useful learning signals for policy optimization in zero-variance settings. Higher semantic entropy is associated with more diverse solution distributions (Chen et al., 2025a), indicating multiple plausible reasoning paths, whereas lower entropy reflects more consistent outputs and stronger model confidence. In our implementation, we cluster rollout embeddings derived from vLLM hidden states with K-means using a fixed $K = 2$; we study the sensitivity to K in Section 5.4.

Based on semantic entropy, LZPO assigns response-level advantages to zero-variance prompts. For *all_correct* prompts, it assigns a weak positive advantage proportional to semantic entropy, which preserves performance on easier problems while giving responses with higher semantic entropy greater relative advantage within the group and encouraging exploration of diverse solution paths. The treatment of *all_wrong* prompts is detailed in Section 4.2.

4.2 Reflection-guided advantage design for *all_wrong* prompts

Although semantic entropy provides a principled measure for reactivating zero-variance prompts, it alone is insufficient. In particular, it cannot distinguish erroneous responses

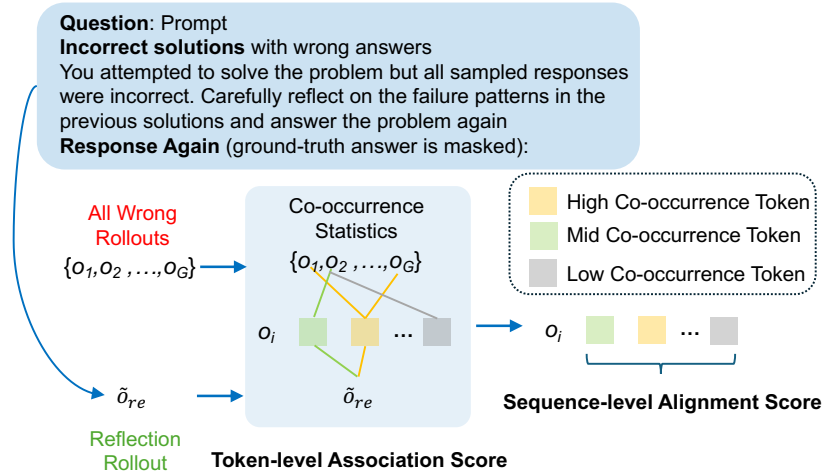


Figure 3: Reflection prompt template and response association via reflection on all-wrong prompts.

with constructive exploratory value from purely noisy failures, nor can it reliably identify which incorrect responses contain latent reasoning paths toward the correct solution.

Inspired by prior work (Sun et al., 2025; Li et al., 2025), we introduce an *error reflection* mechanism to refine advantage assignment and provide more targeted guidance for exploration on hard problems. The core idea is to let the model summarize and reflect on erroneous responses, generate a revised response that approximates the correct reasoning direction under the current policy, and then evaluate the alignment between each erroneous response and this direction.

Reflection-based guidance for *all_wrong* prompts. For an *all_wrong* prompt, we adopt the reflection procedure illustrated in Figure 3. When all rollouts of a prompt are incorrect, we reconstruct the erroneous responses using the input prompt template adopted during training, with the ground-truth answer explicitly masked. This encourages the model to summarize the observed mistakes and re-attempt the problem using only its internal reasoning. The process yields a reflection response $\tilde{o}_{re}$, which provides a weak directional signal under the current policy. We use the reflection response only to rank the relative quality of the original erroneous responses. It is neither treated as verified ground truth nor assigned directly as a positive training target.

Token-level association for reflection alignment. For an *all_wrong* prompt, we measure how each erroneous response aligns with the reflection response $\tilde{o}_{re}$, which serves as a weak proxy for a more promising reasoning direction under the current policy. Concretely, for each token $o_{i,j}$ in the erroneous rollouts, we compute a token-level association score based on its co-occurrence with $\tilde{o}_{re}$, using a combination of Fisher’s exact test and information gain:

$$S(o_{i,j}) = 0.5 F(o_{i,j}) + 0.5 IG(o_{i,j}). \quad (8)$$

Detailed definitions are provided in **Appendix A**.

We then aggregate token-level scores into a sequence-level alignment score $\bar{S}(o_i) = \frac{1}{|o_i|} \sum_j S(o_{i,j})$, and further average them within each semantic cluster to obtain $\bar{S}(C_i)$, which guides sequence-level advantage construction. We use token-level statistics rather than direct sequence-level similarity because reflection provides only weak directional supervision, while erroneous rollouts often contain partially correct reasoning steps that are better captured locally.

Final advantage design. When the group-level reward variance collapses to zero, LZPO activates the semantic entropy and reflection-guided advantage mechanism; otherwise, it reduces to standard GRPO. Let $H_{SE}^{\max} := \max_{C \in \mathcal{C}} H_{SE}(C)$ denote the maximum cluster-level

semantic entropy in the group, and let $C_{\text{worst}} := \arg \min_{C \in \mathcal{C}} \bar{S}(C)$ denote the semantic cluster with the lowest alignment score relative to the reflection response. The advantage is defined as

$$\hat{A}_i^{\text{LZPO}} = \begin{cases} \alpha H_{\text{SE}}(C_i), & \text{all_correct,} \\ -\alpha (H_{\text{SE}}^{\text{max}} - H_{\text{SE}}(C_i)), & \text{all_wrong} \wedge C_i \neq C_{\text{worst}}, \\ -\alpha H_{\text{SE}}^{\text{max}}, & \text{all_wrong} \wedge C_i = C_{\text{worst}}, \\ \hat{A}_{i,t}^{\text{GRPO}}, & \text{otherwise.} \end{cases} \quad (9)$$

where α denotes the scaling-factor hyperparameter. For *all_wrong* prompts, semantic clusters are ranked by their aggregated alignment scores $\bar{S}(C_i)$ with the reflection response. The alignment score is used only for relative ranking among erroneous clusters and never assigns a positive advantage to an incorrect response. The cluster C_{worst} receives the stronger negative advantage $-\alpha H_{\text{SE}}^{\text{max}}$, while all remaining clusters receive the softened penalty $-\alpha (H_{\text{SE}}^{\text{max}} - H_{\text{SE}}(C_i))$ so that potentially informative exploration is penalized less severely.

5 Experiment

5.1 Experimental settings

Benchmarks and parameter details. We conduct experiments across two model scales, DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025), to evaluate the effectiveness and scalability of the proposed method. All models are trained on the DAPO-Math-17K dataset (Yu et al., 2025). During training, we use a rollout size of $N = 8$, a sampling temperature of 1.0, a maximum response length of 8192 tokens, and a fixed learning rate of $\eta = 1 \times 10^{-6}$. For fair comparison, all baselines are reproduced under identical hyperparameter settings on the VeRL platform (Sheng et al., 2025), with GRPO as the base post-training algorithm. We evaluate performance on six mathematical reasoning benchmarks: AIME 2024/2025 (Codeforces, 2024), AMC 2023 (Mathematical Association of America, 2023), Minerva (Lewkowycz et al., 2022), MATH500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024). We report Avg@32 for all benchmarks except MATH-500 and OlympiadBench, for which Avg@4 is used. To assess generalization beyond mathematical reasoning, we additionally evaluate LiveCodeBench (Jain et al., 2024) and IFBench-P (Pyatkin et al., 2025), using Avg@4 and a maximum response length of 25,000 tokens for both benchmarks. We also evaluate Qwen3-8B-Base (Yang et al., 2025) to examine whether the gains extend beyond the DeepSeek-R1-Distill-Qwen model family. Additional experimental details are provided in Appendix E.

Baselines. We compare our method with several representative baselines, including GRPO, GRPO with dynamic sampling that filters zero-variance prompts during training (Yu et al., 2025), GRESO (Zheng et al., 2025), ZVP (Le et al., 2025), and SEED-GRPO (Chen et al., 2025a). We also compare with reflection-based methods that directly update policy gradients using reflected prompts (Li et al., 2025). These baselines cover data filtering, token-level advantage shaping, semantic-entropy-based update scaling, and reflection-guided optimization.

5.2 Main results

As shown in Table 1, LZPO consistently outperforms GRPO on all six mathematical benchmarks, with gains of up to **3.3 points**, particularly on challenging and contamination-free benchmarks such as AIME24 and AIME25 (Wu et al., 2025b). The average gains over GRPO are 1.2 points for the 1.5B model and 2.0 points for the 7B model. LZPO also outperforms SEED-GRPO on five of the six benchmarks and improves the average score from 62.5 to 63.5 on the 7B model.

Beyond mathematical reasoning, LZPO improves over GRPO from 23.5 to 26.7 on LiveCodeBench and from 11.7 to 12.9 on IFBench-P for the 1.5B model. The corresponding scores improve from 49.1 to 52.4 and from 18.9 to 19.5 for the 7B model. On Qwen3-8B-Base, LZPO improves the average score from 47.3 for GRPO and 52.5 for ZVP to 55.4. Full cross-task and cross-model results are reported in Appendix C.

Method	AIME24 Avg@32	AMC23 Avg@32	Olympiad Avg@4	AIME25 Avg@32	MATH-500 Avg@4	Minerva Avg@32	Average
<i>DeepSeek-R1-Distill-Qwen-1.5B</i>							
Vanilla	29.9	63.8	45.9	22.4	84.9	18.4	44.2
GRPO	31.9	76.9	56.0	23.9	86.6	21.5	49.5
GRPO+DS	<u>34.4</u>	77.4	<u>56.3</u>	25.8	87.2	<u>22.1</u>	<u>50.5</u>
GRESO	35.0	<u>78.2</u>	<u>56.3</u>	25.7	87.2	21.5	50.7
ScRPO	34.2	68.5	52.0	26.3	84.1	22.0	47.9
ZVP	32.9	77.1	54.5	24.3	84.1	19.9	48.8
+LZPO (Ours)	34.0	78.4	56.4	<u>25.9</u>	<u>86.7</u>	22.5	50.7
Δ (vs. GRPO)	+2.1	+1.5	+0.4	+2.0	+0.1	+1.0	+1.2
<i>DeepSeek-R1-Distill-Qwen-7B</i>							
Vanilla	52.3	82.3	64.0	39.1	90.7	26.7	59.2
GRPO	54.6	87.8	64.7	40.9	93.0	28.1	61.5
GRPO+DS	52.7	88.7	<u>66.6</u>	34.1	93.4	28.4	60.6
GRESO	52.7	89.5	<u>65.3</u>	35.2	92.7	28.8	60.7
ScRPO	<u>57.5</u>	83.5	65.1	42.2	91.6	<u>29.1</u>	61.5
ZVP	53.1	84.0	62.4	38.2	90.8	27.1	59.3
SEED-GRPO	56.5	89.2	66.4	40.9	92.3	29.7	<u>62.5</u>
+LZPO (Ours)	57.9	90.6	68.6	<u>41.7</u>	<u>93.3</u>	29.0	63.5
Δ (vs. GRPO)	+3.3	+2.8	+3.9	+0.8	+0.3	+0.9	+2.0

Table 1: Experimental results on six mathematical reasoning benchmarks comparing LZPO with multiple baselines, including ZVP and SEED-GRPO. The best and second-best results under each setting are highlighted in **bold** and underline, respectively.

LZPO incurs conditional overhead because reflection requires an additional generation stage for *all_wrong* prompt groups. Relative to GRPO, the end-to-end training time increases by 33% for the 1.5B model and by 24% for the 7B model. On the 7B model, the relative wall-clock costs of GRPO, ZVP, LZPO, and GRPO+DS are 100%, 100%, 124%, and 150%, respectively. This comparison includes the full cost of reflection generation.

5.3 Training dynamics and exploratory behavior

As shown in Figure 4, LZPO exhibits faster entropy decay than GRPO during the first 300 training steps for the 7B model, indicating earlier convergence and more effective use of informative learning signals. This trend is consistent with the findings of ZVP (Le et al., 2025), suggesting that **zero-variance prompts provide additional learning signals**. In contrast, for the 1.5B model, entropy decreases rapidly at the early stage and then stabilizes because of the model’s limited initial capacity. As training proceeds, LZPO steadily reduces entropy and maintains stable training dynamics, whereas GRPO repeatedly suffers from *entropy collapse* in the later stage of 1.5B training, accompanied by chaotic generations and error loops (see Appendix F). Consistent with these dynamics, LZPO outperforms GRPO at the final stage of training, and the performance gap is larger at the 7B scale.

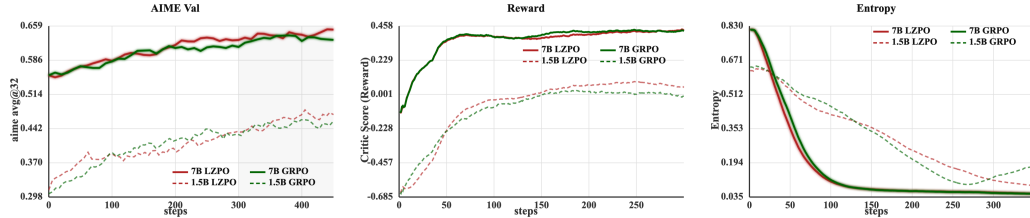


Figure 4: Training dynamics comparison between LZPO and GRPO on 7B and 1.5B models. **Left:** AIME24 Accuracy; **Middle:** Reward Score; **Right:** Generation Entropy.

We attribute this behavior to the evolving composition of effective training signals. Early in training, informative samples are abundant, and both methods benefit from them. For the 1.5B model, limited initial reasoning ability causes *all_wrong* samples to dominate, reaching

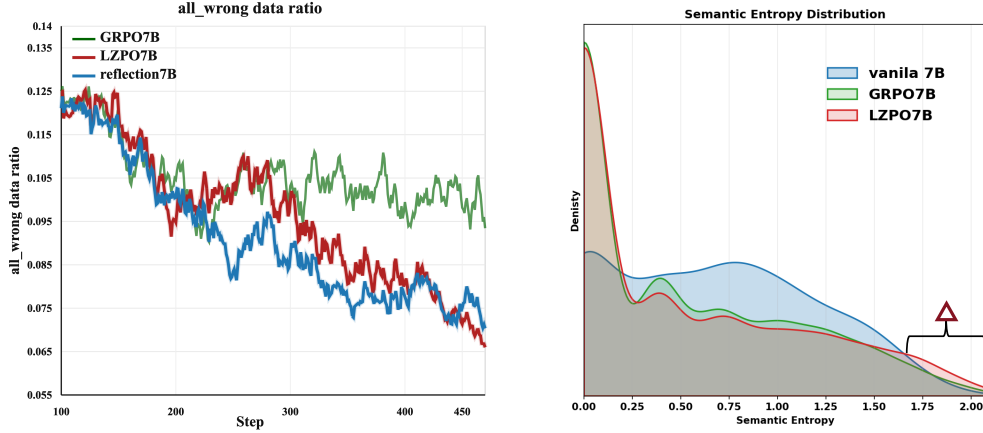


Figure 5: **Left:** evolution of the *all_wrong* prompt ratio during training on the 7B model, comparing GRPO, LZPO, and the reflection-based method. **Right:** semantic entropy distribution of model rollouts at the final training stage.

nearly 60% in Figure 1 Right. Under this condition, LZPO can already extract useful signals from zero-variance prompts and thus gains an early advantage. At later training stages, although the proportion of effective samples drops to around 60%, the fraction of *all_wrong* samples remains largely unchanged, whereas the fraction of *all_correct* samples continues to increase, indicating performance saturation. Under this regime, GRPO is more vulnerable to degenerate behaviors such as repetitive reasoning and semantic drift, especially for smaller models. In contrast, LZPO maintains stable training dynamics and stronger logical reasoning. For the 7B model, as the number of non-zero-variance samples rapidly decreases, LZPO continues to provide more informative learning signals than GRPO, leading to a larger accuracy gap at the later stage of training. We further verify the exploratory effect at the semantic level through the semantic-entropy distribution in Figure 5 Right, computed from training data sampled over 100 consecutive steps at the final stage and comparing Vanilla-7B, GRPO-7B, and LZPO-7B. In particular, LZPO maintains higher density than GRPO in the high normalized semantic-entropy range of $[0.8, 1.0]$, showing that it better preserves semantic diversity within rollout groups and thus encourages exploration. Qualitative analysis in Appendix F further shows that LZPO-trained models better reconstruct core constraints, organize counting logic, and explore more diverse solution paths during rollout, whereas GRPO often degenerates into shallow trial-and-error with repetitive reasoning and semantic drift.

High-difficulty reasoning. LZPO shows stronger advantages and greater exploratory capability on difficult mathematical reasoning tasks. Compared with GRPO, LZPO and hard-prompt reflection methods that exploit *all_wrong* prompts continue to improve performance at the final stage of training. Under a batch size of 256, the proportion of *all_wrong* prompts decreases by up to 3%, as shown in Figure 5 Left, indicating that the model continues to push the upper bound of reasoning capability. Compared with data-filtering methods such as GRPO+DS and GRESO, LZPO achieves substantially larger gains on more challenging benchmarks. Despite the conditional overhead introduced by reflection generation, the 7B model achieves average gains of 5.2 and 6.5 points on AIME24 and AIME25, respectively. Overall, these results demonstrate that LZPO yields superior performance and training stability across model scales by making more effective use of zero-variance signals.

5.4 Ablation study

To disentangle the contributions of the proposed designs, we conduct systematic ablations on the 7B model and organize the analysis into five parts, as summarized in Table 2.

Reactivating zero gradients and the necessity of both mechanisms. We first examine whether the gains arise merely from reintroducing non-zero gradients on zero-variance prompts. The simple non-zero-advantage baseline assigns fixed advantages of +0.3 and -0.3 to *all_correct* and *all_wrong* prompts, respectively. The reflection-only and semantic-entropy-only variants outperform this baseline by 1.1 and 0.5 points on average, showing

Method	AIME24 @32	AMC @32	Olymp. @4	AIME25 @32	MATH500 @4	Minerva @32	Avg
<i>Ablation Study on DeepSeek-R1-Distill-Qwen-7B</i>							
<i>(I) Reactivating zero gradients and the necessity of both mechanisms</i>							
Reflection only	58.3	90.2	67.7	42.2	93.0	29.3	63.5
Semantic entropy only	57.1	91.0	67.1	39.4	93.7	28.2	62.8
Simple non-zero advantage	56.7	90.9	66.4	39.2	93.1	27.4	62.3
<i>(II) Effect of entropy-based advantage scaling α</i>							
$\alpha = 0.1$	56.3	90.2	67.1	40.3	93.6	29.1	62.8
$\alpha = 0.2$	57.4	90.9	67.7	41.0	93.2	27.8	63.0
$\alpha = 0.4$	57.4	91.1	67.9	39.1	<u>93.9</u>	<u>29.3</u>	63.1
<i>(III) Robustness of semantic clustering and compatibility with advanced components</i>							
$K = 3$	57.4	91.2	67.8	40.5	93.3	28.3	63.1
$K = 4$	56.5	90.0	66.1	38.9	91.8	26.6	61.7
LZPO ($K = 2, \alpha = 0.3$)	57.9	90.6	68.6	41.7	93.3	29.0	63.5
LZPO + Clip-Higher	<u>58.7</u>	91.2	68.0	43.3	94.1	27.9	<u>64.0</u>
<i>(IV) Equal-budget control for reflection</i>							
Extra sampling ($N = 12$)	54.7	88.3	65.4	41.1	93.7	27.6	61.8
<i>(V) Larger rollout groups</i>							
$N = 32, K = 3$	58.5	<u>91.4</u>	68.4	41.7	94.1	29.5	63.9
$N = 32, K = 4$	59.3	92.2	68.8	<u>42.4</u>	93.5	28.9	64.2

Table 2: Ablation study on DeepSeek-R1-Distill-Qwen-7B. Group (I) compares the two proposed components with simple non-zero advantage assignment. Group (II) studies the scaling factor α . Group (III) studies clustering granularity with the default $N = 8$ configuration and compatibility with Clip-Higher. Group (IV) replaces reflection with four additional standard rollouts under a comparable generation budget. Group (V) evaluates finer-grained clustering with larger rollout groups. All metrics are reported as Avg@k.

that the gains arise from structured advantage construction rather than gradient reactivation alone. The two mechanisms are also complementary across benchmarks. Reflection provides larger gains on AIME24, AIME25, and Minerva, where correcting unsuccessful reasoning directions is particularly important. Semantic entropy provides larger gains on AMC23 and MATH-500, where the baseline performance is higher and additional successful reasoning paths can be exploited. Their improvements are comparable on OlympiadBench.

We further evaluate error reflection and token-level associations for semantic relevance. After error reflection, nearly 35% of revised responses are corrected, and many responses previously trapped in error loops are repaired. Using an auxiliary GPT-5.5 judge, we find that erroneous responses with higher token-level relevance to the reflected rollout tend to receive higher scores. This suggests that responses more aligned with the reflected rollout are more likely to contain useful reasoning signals and thus deserve milder penalties. More details are provided in [Appendix B](#).

Equal-budget reflection control and larger rollout groups. The reflection-only variant obtains an average score of 63.5, compared with 61.8 for the equal-budget extra-sampling control, and performs better on five of the six benchmarks. This result indicates that targeted revision provides a more informative signal than additional independent sampling. Under the default $N = 8$ setting, $K = 4$ leaves approximately two responses per cluster and reduces the average score to 61.7. With $N = 32$, the $K = 3$ and $K = 4$ configurations achieve 63.9 and 64.2, respectively, suggesting that finer-grained semantic partitions become more reliable when more responses are available. These larger-group results are not directly comparable with the default configuration because increasing N also changes rollout cost, response diversity, and the frequency of zero-variance groups.

Effect of entropy-based advantage scaling. We next study the effect of the scaling factor α . Because semantic entropy is bounded, inappropriate scaling can produce either weak learning signals or unstable optimization. Empirically, $\alpha = 0.3$ provides the best trade-off between performance and stability on the 7B model. Smaller values lead to insufficient

learning signals, whereas larger values produce overly aggressive updates and can suppress the contribution of non-zero-variance samples. These results indicate that the effect of advantage shaping depends not only on the presence of non-zero signals but also on their calibration.

Robustness of semantic clustering and compatibility with advanced components. We finally analyze the role of clustering and the compatibility of LZPO with existing optimization strategies. In our implementation, each prompt has $N = 8$ rollouts, and semantic entropy is computed by applying K-means to rollout embeddings with a fixed small K , using $K = 2$ by default. This choice improves stability and interpretability under small-sample clustering, whereas larger K values are more prone to K-means noise and unstable partitions. As shown in Table 2, the performance difference between $K = 2$ and $K = 3$ is small, indicating limited sensitivity to moderate changes in K once variance effects are controlled. We further find that LZPO is compatible with existing methods because it operates at the level of advantage shaping. For example, combining LZPO with Clip-Higher yields additional gains. This result suggests that the semantic-entropy and reflection mechanisms in LZPO are complementary not only to standard clipping-based optimization, but also to more advanced post-training algorithms.

6 Conclusion

We introduce LZPO to recover learning signals from zero-variance prompt groups that provide no effective advantage under GRPO. LZPO applies semantic-entropy-guided advantages to *all.correct* groups and reflection-guided relative penalties to *all.wrong* groups. Experiments across mathematical reasoning, code generation, and verifiable instruction following demonstrate consistent improvements over GRPO and related baselines, while the results on Qwen3-8B-Base show that the gains extend beyond a single model family. Component ablations further show that semantic entropy and reflection provide complementary benefits, and the equal-budget control confirms that the gain from reflection cannot be explained by additional sampling alone. These improvements require conditional reflection generation, whose end-to-end cost is included in our analysis.

Limitations

Although our evaluation covers mathematical reasoning, code generation, and verifiable instruction following, the policies are primarily trained on mathematical data. The cross-task results therefore measure transfer rather than domain-specific RL training. The model coverage remains limited to two model families and three model configurations. Under the default $G = 8$ and $K = 2$ setting, semantic entropy provides only a coarse-grained estimate of semantic-mode imbalance. Reflection can remain incorrect and may introduce length and style biases; in our sampled analysis, reflection responses contain 577 more tokens than ordinary rollouts on average. Reflection generation also increases end-to-end training time by 33% for the 1.5B model and by 24% for the 7B model. Finally, LZPO relies on fixed clustering and advantage-scaling hyperparameters, motivating future work on adaptive clustering and advantage calibration.

Ethics Statement

LLMs were used for language polishing and, where explicitly reported, for auxiliary analysis of reflection responses. LLM-based judgments do not determine verifier-based benchmark scores or training rewards.

Acknowledgement

This work was supported by the Open Research Fund of the Shanghai Key Laboratory of Quantum City Spatial Intelligence Innovation (Grant No. KF-2026-01).

References

Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1): 131–153, 1992.

- Hongye Cao, Zhixin Bai, Ziyue Peng, Boyan Wang, Tianpei Yang, Jing Huo, Yuyao Zhang, and Yang Gao. Efficient reinforcement learning with semantic and token entropy for llm reasoning, 2025. URL <https://arxiv.org/abs/2512.04359>.
- Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization, 2025a. URL <https://arxiv.org/abs/2505.12346>.
- Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@k training for adaptively balancing exploration and exploitation of large reasoning models, 2025b. URL <https://arxiv.org/abs/2508.10751>.
- Codeforces. American Invitational Mathematics Examination (AIME) 2024. <https://codeforces.com/>, 2024. Online.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL <https://arxiv.org/abs/2505.22617>.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Jia Deng, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, and Ji-Rong Wen. Decomposing the entropy-performance exchange: The missing keys to unlocking effective reinforcement learning, 2025. URL <https://arxiv.org/abs/2508.02260>.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong

- Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL <https://arxiv.org/abs/2402.14008>.
- 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, 2021. URL <https://arxiv.org/abs/2103.03874>.
- Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward, 2025. URL <https://arxiv.org/abs/2510.03222>.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL <https://arxiv.org/abs/2403.07974>.
- Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL <https://arxiv.org/abs/2406.15927>.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023. URL <https://arxiv.org/abs/2302.09664>.
- Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in llm reinforcement learning via entropy-guided advantage shaping, 2025. URL <https://arxiv.org/abs/2509.21880>.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL <https://arxiv.org/abs/2206.14858>.
- Lianrui Li, Dakuan Lu, Jiawei Shao, Chi Zhang, and Xuelong Li. Scrpo: From errors to insights, 2025. URL <https://arxiv.org/abs/2511.06065>.
- Mengqi Liao, Xiangyu Xi, Ruinian Chen, Jia Leng, Yangen Hu, Ke Zeng, Shuai Liu, and Huaiyu Wan. Enhancing efficiency and exploration in reinforcement learning for llms, 2025. URL <https://arxiv.org/abs/2505.18573>.
- Mathematical Association of America. American Mathematics Competitions (AMC). <https://maa.org/>, 2023. Online.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sult, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman,

- Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiye Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL <https://arxiv.org/abs/2412.16720>.
- Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. URL <https://arxiv.org/abs/2507.02833>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys ’25*, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL <http://dx.doi.org/10.1145/3689031.3696075>.
- Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning, 2025. URL <https://arxiv.org/abs/2509.20712>.
- Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. Ktae: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning, 2025. URL <https://arxiv.org/abs/2505.16826>.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL <https://arxiv.org/abs/2506.01939>.

- Hao Wu and Wei Liu. Gcpo: When contrast fails, go gold, 2025. URL <https://arxiv.org/abs/2510.07790>.
- Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation for entropy-safe reasoning, 2025a. URL <https://arxiv.org/abs/2509.22611>.
- Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination, 2025b. URL <https://arxiv.org/abs/2507.10532>.
- Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning, 2025. URL <https://arxiv.org/abs/2510.10649>.
- Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework under non-linear rl objectives, 2025. URL <https://arxiv.org/abs/2510.04996>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.
- Yu Yue, Yufeng Yuan, Qiyang Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, Tiantian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks, 2025. URL <https://arxiv.org/abs/2504.05118>.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization, 2025. URL <https://arxiv.org/abs/2504.05812>.
- Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts, 2025. URL <https://arxiv.org/abs/2506.02177>.

A Token association

For an *all_wrong* prompt, we consider a set of erroneous rollouts $x_F = \{o_1, \dots, o_G\}$ together with a reflection response $\tilde{o}_{re}$, which serves as a proxy for the approximate correct reasoning direction under the current policy. We define the proxy-correct set as $x_T = \{\tilde{o}_{re}\}$ and let $N = |x_T| + |x_F| = G + 1$.

Hyperparameters	DeepSeek-R1-Distill- Qwen-1.5B	DeepSeek-R1-Distill- Qwen-7B
Epochs	6	6
rollout size	8	8
temperature	1.0	1.0
top-p	1.0	1.0
top-k	-1	-1
learning rate	1×10^{-6}	1×10^{-6}
max response length	8192	8192
global batch size	256	256
PPO mini-batch size	128	128
clip ratio	(0.2, 0.2)	(0.2, 0.2)
temperature(Eval)	0.6	0.6
top-p(Eval)	1.0	1.0
max response length(Eval)	8192	8192

Table 3: Training hyperparameters used in our experiments.

For each token $o_{i,j}$ appearing in the erroneous responses x_F , we summarize its co-occurrence statistics with the reflected reasoning direction using a 2×2 contingency table (Sun et al., 2025), which records whether the token appears at least once in proxy-correct versus erroneous rollouts. Concretely, we define four occurrence counts as follows: $a(o_{i,j}) = \sum_{x \in x_T} \mathbf{1}(o_{i,j} \in x)$ counts the number of proxy-correct rollouts containing $o_{i,j}$, $c(o_{i,j}) = |x_T| - a(o_{i,j})$ counts those that do not; analogously, $b(o_{i,j}) = \sum_{x \in x_F} \mathbf{1}(o_{i,j} \in x)$ counts the number of erroneous rollouts containing $o_{i,j}$, and $d(o_{i,j}) = |x_F| - b(o_{i,j})$ counts those that do not. By construction, these counts satisfy $a(o_{i,j}) + b(o_{i,j}) + c(o_{i,j}) + d(o_{i,j}) = N$. This contingency table characterizes the statistical alignment between token usage and the reflection-indicated reasoning direction.

Association strength. Based on the contingency table, we quantify the association between token $o_{i,j}$ and the reflected reasoning direction by combining a statistical significance measure with an information-theoretic measure. We apply Fisher’s exact test (Agresti, 1992; Sun et al., 2025) under the null hypothesis that the occurrence of token $o_{i,j}$ is independent of rollout correctness. The resulting p -value measures the likelihood of observing the contingency table (or a more extreme one) under independence.

Information gain. To complement statistical significance with an information-theoretic perspective, we compute the information gain (IG) between token occurrence and rollout correctness (Sun et al., 2025). Let Y denote a binary variable indicating whether a rollout aligns with the reflected reasoning direction, and let $X_{o_{i,j}}$ indicate whether token $o_{i,j}$ appears. The information gain is defined as

$$\text{IG}(o_{i,j}) = H(Y) - H(Y \mid X_{o_{i,j}}). \quad (10)$$

which measures how much knowing the presence of $o_{i,j}$ reduces uncertainty about rollout alignment.

Combined association score. We combine the two complementary signals into a single token-level association score (Sun et al., 2025),

$$S(o_{i,j}) = 0.5 F(o_{i,j}) + 0.5 \text{IG}(o_{i,j}). \quad (11)$$

where $F(o_{i,j})$ denotes the transformed Fisher significance score. As discussed in the main text, this token-level association score is *not* used for direct token-level policy updates.

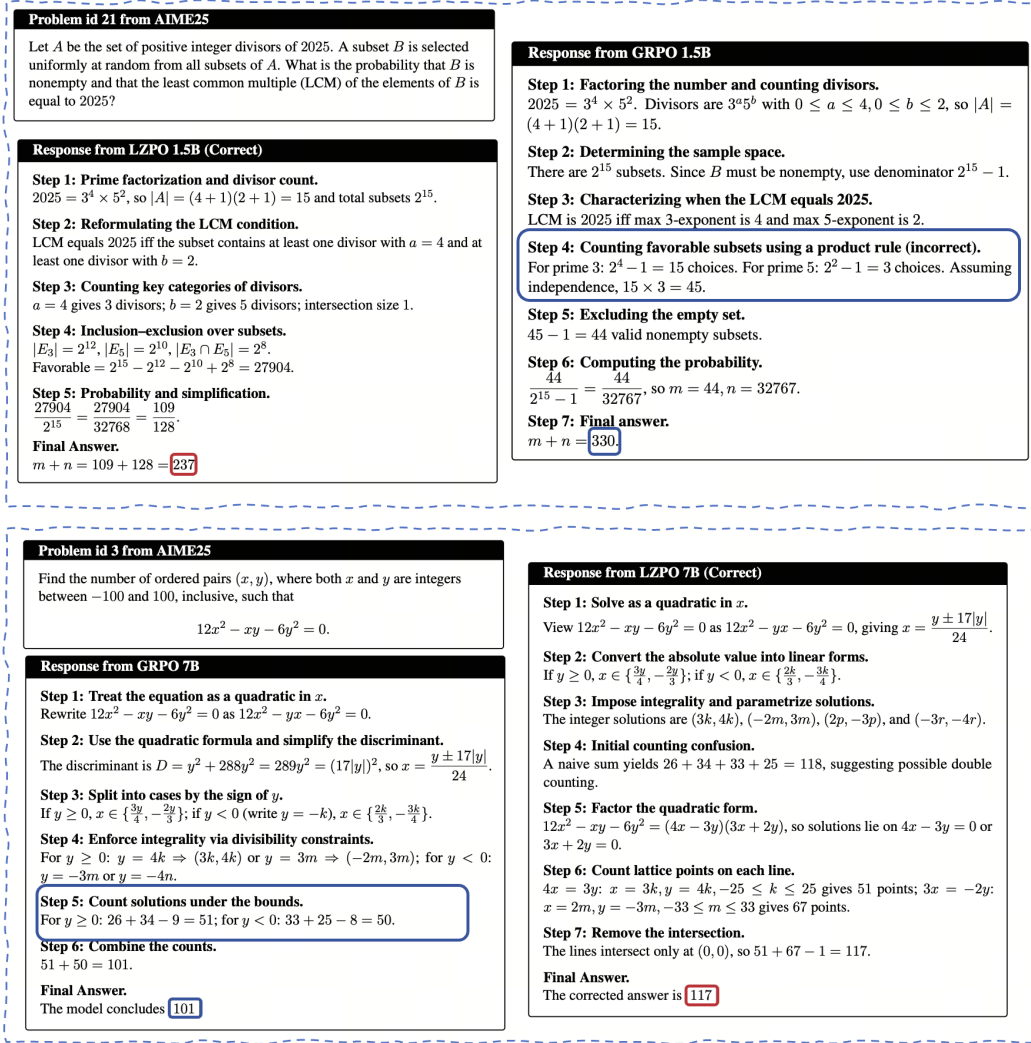


Figure 6: Qualitative comparison of GRPO and LZPO responses on AIME25. Top: 1.5B scale. Bottom: 7B scale.

Instead, it is aggregated at the response level to assess alignment with the reflection-indicated reasoning direction, and is used solely to modulate response-level advantages in the *all_wrong* setting.

B Reflection statistics

We distinguish offline reflection accuracy from accuracy during online RL training. Among 1,000 offline failures sampled from six mathematical benchmarks, reflection corrects approximately 35% of the responses. During online training of DeepSeek-R1-Distill-Qwen-7B with a batch size of 256, approximately 11% of prompt groups trigger reflection, and 32% of the reflection responses generated between Steps 100 and 500 are verified as correct. The online correction rate remains broadly stable but shows a modest upward trend, reaching a level 4–6 percentage points higher in the later stage than in the earlier stage.

Among 100 successful reflections, the model often identifies local but critical errors, including omitted constraints, incomplete case analysis, and incorrect intermediate calculations. Among 100 unsuccessful reflections, the revised reasoning is usually more complete but can preserve central misunderstandings, particularly in geometric reconstruction, piecewise

reasoning, and global constraint checking. These observations support the use of reflection as weak directional supervision rather than verified ground truth.

We further use a GPT-5.5 judge to assess reasoning consistency and answer quality. Responses that deviate from a successful reflection direction generally receive lower judge scores and stronger negative advantages. When reflection remains incorrect, higher-scoring responses tend to satisfy more problem constraints and remain closer to a valid solution path. This judge-based analysis supports relative ranking but does not justify assigning positive rewards to reflection responses.

Reflection can also introduce length and style biases. Across 200 sampled prompts, reflection responses contain 577 more tokens on average than ordinary rollouts and use self-checking expressions more frequently. Semantic alignment may therefore favor longer responses or responses written in a reflective style.

C Additional evaluation

ZVP reproduction. Because the official implementation of ZVP is not publicly available, we independently reproduce its token-entropy-based advantage mechanism under our training configuration. On the 1.5B model, ZVP reduces generation entropy more rapidly and converges approximately 200 steps earlier than GRPO and LZPO, but it also exhibits unstable generations and partial entropy collapse. This behavior is consistent with the sensitivity of direct token-entropy weighting when incorrect samples account for a large fraction of the training data. In contrast, LZPO constructs response-level signals and uses reflection-guided relative penalties for *all_wrong* groups, resulting in more stable optimization under this setting.

Cross-task evaluation. Table 4 reports results on LiveCodeBench and IFBench-P. All methods are evaluated using Avg@4 with a maximum response length of 25,000 tokens.

Model	Method	LiveCodeBench Avg@4	IFBench-P Avg@4
1.5B	GRPO	23.5	11.7
	GRPO+DS	<u>26.0</u>	<u>12.6</u>
	ZVP	23.8	12.0
	LZPO	26.7	12.9
7B	GRPO	49.1	18.9
	GRPO+DS	<u>51.3</u>	<u>19.4</u>
	ZVP	50.3	19.0
	LZPO	52.4	19.5

Table 4: Cross-task evaluation of the DeepSeek-R1-Distill-Qwen models on code generation and verifiable instruction following. The best and second-best results for each model scale are highlighted in **bold** and underline, respectively.

Cross-model evaluation. Table 5 reports results on Qwen3-8B-Base. For ZVP, we use the result reported under the corresponding Qwen3-8B setting because its official implementation is unavailable. LZPO improves the average score by 8.1 points over GRPO and by 2.9 points over ZVP.

Qwen3-8B-Base							
Method	AIME24	AMC23	Olympiad	AIME25	MATH-500	LiveCodeBench	Average
GRPO	28.3	73.8	49.6	20.0	83.0	29.2	47.3
ZVP	<u>33.3</u>	<u>78.8</u>	<u>58.2</u>	<u>24.6</u>	<u>89.7</u>	<u>30.1</u>	<u>52.5</u>
LZPO	34.0	90.4	60.0	26.2	90.4	31.1	55.4

Table 5: Cross-model evaluation on Qwen3-8B-Base. The best and second-best results are highlighted in **bold** and underline, respectively.

D Efficiency analysis

Reflection generation dominates the additional rollout-stage cost of LZPO, while semantic representation, entropy computation, and clustering contribute only a small fraction.

Model	Reflection	Representation and entropy	K-means
1.5B	98%	$\approx 0\%$	2%
7B	95%	$\approx 0\%$	5%

Table 6: Composition of the additional rollout-stage cost of LZPO for the DeepSeek-R1-Distill-Qwen models. Percentages describe the distribution within the additional cost rather than fractions of total training time.

E Details on experimental design

Experimental parameter settings. To ensure reproducibility and clarity of the experimental setup, we report all hyperparameters used in our experiments in Table 3. Any hyperparameters not explicitly listed follow the default configuration of the VeRL (Sheng et al., 2025) framework. All methods are trained for 1000 steps using the same hardware, base rollout configuration, and optimization hyperparameters. LZPO additionally performs conditional reflection generation for *all_wrong* prompt groups, and this computation is fully included in the reported end-to-end training time. All experiments are conducted on 8 NVIDIA A100 GPUs, and to ensure the reliability of the results given the stochastic nature of reinforcement learning, each experiment is repeated three times.

F Case study

Figure 6 compares AIME25 responses from the 1.5B and 7B configurations of DeepSeek-R1-Distill-Qwen trained with GRPO and LZPO. GRPO often commits prematurely to fragile heuristics and overlooks key constraints, causing systematic counting errors. In contrast, LZPO more clearly reconstructs the constraints, organizes the counting logic, and handles overlaps and exclusions.

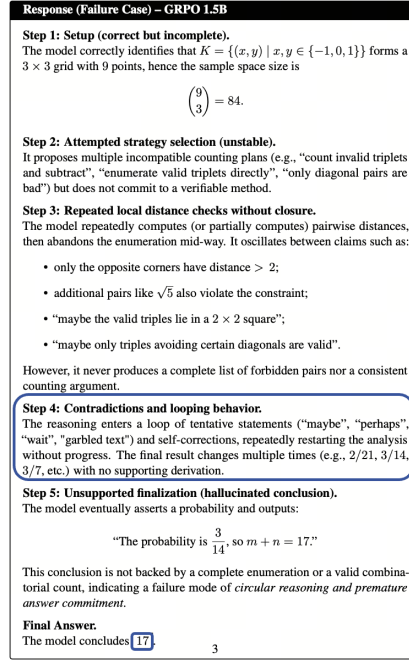


Figure 7: A failure case of GRPO on DeepSeek-R1-Distill-Qwen-1.5B during late-stage training.

Figure 7 shows a late-stage GRPO failure in which the model cycles among incompatible strategies. LZPO instead uses semantic entropy and reflection to maintain coherent reasoning.