Beyond High-Entropy Exploration: Correctness-Aware Low-Entropy Segment-Based Advantage Shaping for Reasoning LLMs

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has become a central approach for improving the reasoning ability of large language models. Recent work studies RLVR through token entropy, arguing that high-entropy tokens drive exploration and should receive stronger updates. However, they overlook the fact that most of a reasoning trajectory consists of low-entropy segments that encode stable and reusable structural patterns. Through qualitative and quantitative analyses, we find that the overlap of low-entropy segments across correct responses strongly correlates with model accuracy, while overlaps involving incorrect responses exhibit stable but unproductive patterns. Motivated by these findings, we propose LESS, a correctness-aware reinforcement framework that performs fine-grained advantage modulation over low-entropy segments. LESS amplifies segments unique to correct responses, suppresses those unique to incorrect ones, and neutralizes segments shared by both, while preserving high-entropy exploration in the underlying RL algorithm. Instantiated on top of GRPO and GSPO, LESS not only improves accuracy over strong RL baselines across three backbones and six math benchmarks, but also achieves stronger robustness of the performance floor.

1 Introduction

The reasoning capability of Large Language Models (LLMs) plays a central role in tasks such as mathematics (DeepSeek-AI et al., 2025; Chen et al., 2024), programming (5 Team et al., 2025; Wei et al., 2025b; Da et al., 2025), science problem-solving (M2 Team et al., 2025; Sellergren et al., 2025; Jing et al., 2025), and legal analysis (Zhang et al., 2025a,b). Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as an effective approach for improving reasoning reliability, where the correctness of the final answer is used as a reward signal to update the model. Representative RLVR methods (Shao et al., 2024; Yu et al., 2025) typically apply policy updates uniformly across all tokens in a generated sequence.

Recent studies have argued that different parts of a reasoning sequence contribute differently to the final outcome through the lens of token entropy, and that RLVR training should take this into account. They observe that high-entropy tokens often correspond to exploratory reasoning steps, where the model tests alternative solution paths. For example, Cui et al. (2025) show that training encourages the model to explore uncertain reasoning branches; Zhang et al. (2025) encourage diversity in correct attempts by adjusting update strength in high-entropy regions; and Cheng et al. (2025) show that increasing entropy can improve the ability to search for solutions. Most notably, Wang et al. (2025) study demonstrates that only a small subset of tokens with high entropy disproportionately influence reasoning outcomes, suggesting that RL training should focus attention on these regions.

While existing entropy-based approaches have shown promising results, they focus almost entirely on high-entropy tokens, treating these points as the main drivers of exploration in reasoning. This overlooks that most of a reasoning sequence is composed of low-entropy segments (e.g., Wang et al. (2025) treat about 80% of tokens in a response as low-entropy), which form the stable structural scaffold that shapes how the solution is carried out. To examine the role of these low-entropy segments, we conduct both qualitative and quantitative analyses on RLVR-trained models, which is elaborated in §2. From the preliminary studies, we have the following observations. **First**, as shown in Fig 1(a), correct responses share consistent low-entropy segments that reflect coherent and productive reasoning steps, while incorrect responses also display their own repeated low-entropy patterns that represent stable but unproductive reasoning habits. **Second**, cross-dataset evaluation

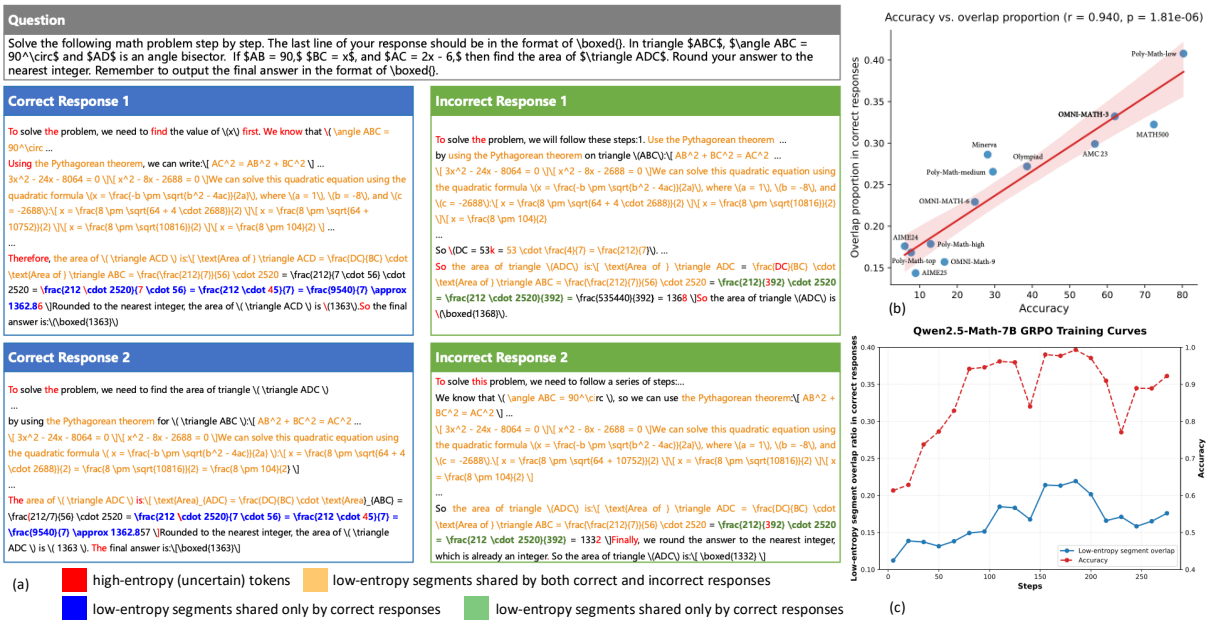


Figure 1: Low-entropy analysis reveals stable reasoning behaviors. **Left**: a case study where correct and incorrect responses exhibit shared and distinct low-entropy segments. **Right-top**: Across math benchmarks, accuracy strongly correlates with low-entropy segment overlap in correct responses. **Right-bottom**: During GRPO training of Qwen2.5-Math-7B, both accuracy and low-entropy overlap rise together, showing that performance gains emerge alongside the stabilization of reasoning patterns.

in Fig 1(b) shows a strong positive correlation between Qwen2.5-Math-7B accuracy and the overlap of low-entropy segments in correct responses, indicating that this relationship holds across tasks and model setups. **Third**, from the training dynamics in Fig 1(c), we observe that the model accuracy and the overlap of low-entropy segments across correct responses increase together, showing that improvements in reasoning ability are accompanied by the consolidation of shared structural patterns. These observations suggest that simply emphasizing high-entropy regions is insufficient, and the treatment of low-entropy segments is directly related to whether useful or harmful reasoning routines are reinforced.

Building on these observations, we introduce **Low-Entropy Segment Shaping (LESS)**, a reinforcement learning with verifiable rewards framework that treats low-entropy structure as an explicit training signal. LESS inserts an entropy-aware segmentation step into the policy update. For each generated solution, it splits the trajectory into high-entropy exploration tokens and contiguous low-entropy segments, and then aggregates how often each segment appears in correct versus incorrect responses within a rollout group. These statistics are used to rescale token-level advantages in a structured way: segments that occur only in correct trajectories receive amplified positive advantages, segments that occur only in incorrect trajectories receive amplified negative advantages, and segments that co-occur in both are neutralized, while high-entropy tokens keep their original RL updates. In this way, LESS strengthens reusable reasoning structure and suppresses repeated failure patterns without harming exploration. The framework is agnostic to the underlying RLVR objective. In this work, we instantiate LESS on top of two widely used multi-sample training methods–Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025). Extensive experiments conducted on six reasoning benchmark demonstrate LESS outperforms popular baseline across almost all tasks and model scales (1.5B, 7B math-tuned, and 7B base). In particular, it yields notable improvements on AIME24/25 and AMC23, where stable multi-step reasoning is essential. Moreover, compared to vanilla GRPO, LESS markedly reduces the worst-case dispersion among sampled responses. This aligns with the core goal of LESS—to reinforce beneficial structural segments and suppress misleading ones—ultimately producing more stable and reliable policy updates.

In summary, our contributions are three-fold: (1) We introduce a segment-level perspective on RLVR that distingsuishes low-entropy segments

by correctness, revealing stable structural patterns in LLM reasoning. (2) We propose LESS, a plug-in algorithm that reweights token-level advantages by segment statistics, amplifying low-entropy segments unique to correct trajectories, suppressing those unique to incorrect ones, and neutralizing shared segments. (3) We show that LESS consistently improves accuracy across six mathematical reasoning benchmarks and three backbones, while also improving robustness under worst@K and reducing variance across sampled rollouts.

2 Preliminary Analysis

We examine low-entropy segments as an indicator of stable reasoning behavior in LLMs and study how these signals relate to model correctness and performance.

We begin by conducting a qualitative experiment to visualize the entropy structure of multiple responses produced for the same question. Specifically, we analyze the token-level entropy patterns of Qwen2.5-Math-7B on the mathematical reasoning dataset (Hendrycks et al., 2021), which enables us to separate responses into stable and unstable regions and to identify the parts of the model's reasoning that remain consistently preserved across different outputs. As shown in Fig. 1 (a), high-entropy tokens (in red) mark unstable regions where the model varies its reasoning, while low-entropy segments reveal stable structures that the model consistently reuses. Within these low-entropy segments, we observe three distinct patterns. (1) Segments shared only by correct responses (in blue) correspond to productive reasoning steps that reliably support the correct solution. (2) Segments shared only by incorrect responses (in green) reflect stable but unproductive reasoning habits, for example, the repeated computation "$\frac{212}{392} \cdot 2520 = \frac{212 \cdot 2520}{392} =$?". (3) Segments shared by both correct and incorrect responses (in orange) capture general reasoning components that are stable but not predictive of correctness—for instance, invoking "the Pythagorean theorem," which provides a common derivation framework but is not the source of the subsequent correct or incorrect calculations. This evidence shows low-entropy segments encode structured reasoning behaviors that differentiate effective and ineffective model responses.

To test whether the qualitative patterns extend beyond a single example, we measure the overlap of low-entropy segments across correct responses for a range of math benchmarks, including AIME24, AIME25, AMC23, MATH500, Minerva, and the Omni-MATH series. As shown in Fig.1 (b), benchmark accuracy is strongly correlated with the degree of low-entropy overlap across correct responses (Pearson $r = 0.94$ and p-value $= 1.81e^{-6}$). Benchmarks with higher accuracy, such as MATH500 and Omni-MATH-3, exhibit clear clustering toward higher overlap ratios, while lower-accuracy benchmarks show weaker consistency in their stable segments. The fitted regression line further highlights this trend, indicating that stronger task performance is associated with more consolidated reasoning structure and greater reuse of stable low-entropy patterns. Similar positive correlations (Appendix A.6) are observed across several other backbones.

We further examine how these patterns evolve during learning. Using GRPO training of Qwen2.5-Math-7B, we track both accuracy and low-entropy segment overlap over training steps. As shown in Fig. 1 (c), the two trajectories rise together throughout training: early stages display low accuracy and fragmented low-entropy structure, while later stages show increasing stability in low-entropy segments alongside improved accuracy. This synchronous growth suggests that the model's reasoning becomes more consistent as training progresses and that stable low-entropy segments emerge as the model acquires more reliable reasoning routines. These results confirm that low-entropy overlap reflects not only final performance but also the developmental trajectory of the model's reasoning behavior. We observe the same co-evolution pattern (Appendix A.5) on Qwen2.5-Math-1.5B and Qwen2.5-7B.

These results show that low-entropy segments provide a reliable signal for understanding and guiding model reasoning. They capture stable computational routines that distinguish correct from incorrect behavior, reflect the degree of structural consistency across benchmarks, and track the development of reasoning stability during learning. These observations suggest low-entropy segments can serve as informative targets for optimization, enabling the model to strengthen productive reasoning routines while suppressing unproductive ones.

3 Methodology

Motivated by these findings, we propose Low-Entropy Segment Shaping (LESS), an RLVR frame-

work that improves reasoning stability by reshaping token-level advantages using statistics of low-entropy segments across sampled responses. LESS is compatible with standard RLVR algorithms and can be used as a plug-in module, and we instantiate it on top of GRPO and GSPO in this work.

3.1 LESS: Low-Entropy Segment Shaping

Given an input question q, the policy generates a group of responses $\mathcal{G} = \{O_1, \ldots, O_G\}$, LESS detects low-entropy segments and shapes corresponding advantages as follows:

Entropy-based segment extraction. For a $O_i = [t_1, \ldots, t_{|O_i|}]$, the entropy of token t_j is

$$\mathcal{H}_{t_j} = -\sum_{v \in V} \pi_{\theta_{\text{old}}}(v \mid x, O_i{<}j) \log \pi_{\theta_{\text{old}}}(v \mid x, O_i{<}j). \tag{1}$$

Following Wang et al. (2025), we compute an entropy threshold τ_i for each response O_i as the h-quantile of its token entropies $\mathcal{H}_{t_j}$. We then treat high-entropy tokens as isolated positions and group consecutive low-entropy tokens into contiguous spans. A minimum length μ is used to filter out trivial low-entropy spans (such as punctuation or very short frequent phrases). This gives three types of entropy-based structures:

$$\begin{aligned}
\mathcal{S}_i^{\text{high}} &= \{t_j \in O_i \mid \mathcal{H}_{t_j} \geq \tau_i\}, \\
\mathcal{S}_i^{\text{frag}} &= \{O_i[a{:}b] \mid b - a + 1 < \mu,\ \forall t_j \in [a, b] :\ \mathcal{H}_{t_j} < \tau_i\}, \\
\mathcal{S}_i^{\text{seg}} &= \{O_i[a{:}b] \mid b - a + 1 \geq \mu,\ \forall t_j \in [a, b] :\ \mathcal{H}_{t_j} < \tau_i\},
\end{aligned} \tag{2}$$

where $\mathcal{S}_i^{\text{high}}$ collects individual high-entropy tokens, $\mathcal{S}_i^{\text{frag}}$ contains short low-entropy fragments that are likely uninformative, and $\mathcal{S}_i^{\text{seg}}$ contains longer low-entropy segments that we regard as structured reasoning candidates. Then, we aggregate how often each $\mathcal{S}_i^{\text{seg}}$ appears in correct versus incorrect responses within a rollout group. Let N_r and N_w denote the number of correct and incorrect responses in $\mathcal{G}$. For a low-entropy segment σ, we count its frequency over the group:

$$\begin{aligned}
n_r(\sigma) &= |\{i \mid correct_i = 1 \wedge \sigma \in \mathcal{S}_i^{\text{seg}}\}|, \\
n_w(\sigma) &= |\{i \mid correct_i = 0 \wedge \sigma \in \mathcal{S}_i^{\text{seg}}\}|.
\end{aligned} \tag{3}$$

Advantage shaping. LESS modifies the advantage assigned to each token t_j in O_i as:

$$\hat{A}_j^i = \begin{cases}
A_i, & t_j \in \mathcal{S}_i^{\text{high}}, \\
A_i / N_r, & t_j \in \mathcal{S}_i^{\text{frag}},\ correct_i = 1, \\
A_i / N_w, & t_j \in \mathcal{S}_i^{\text{frag}},\ correct_i = 0, \\
0, & \sigma_j^i \in \mathcal{S}_i^{\text{seg}},\ n_r > 0,\ n_w > 0, \\
(n_r / N_r) A_i, & \sigma_j^i \in \mathcal{S}_i^{\text{seg}},\ n_r > 0,\ n_w = 0, \\
(n_w / N_w) A_i, & \sigma_j^i \in \mathcal{S}_i^{\text{seg}},\ n_r = 0,\ n_w > 0.
\end{cases} \tag{4}$$

This rule: (i) preserves exploratory high-entropy behavior, (ii) reinforces stable segments unique to correct responses, (iii) penalizes those unique to incorrect responses, (iv) ignores ambiguous segments shared by both groups. The full pseudocode of LESS and its time complexity analysis are given in Appendix A.1.

3.2 Instantiations

LESS is designed as a generic advantage–shaping framework and can be applied to current RLVR methods that computes token- or sequence-level advantages. In this work, we instantiate LESS using the GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025). These methods are particularly suitable for our setting because it (i) generates a group of responses for each query, allowing entropy-based statistics to be computed across samples, and (ii) performs stable clipped-ratio updates that interact well with our advantage shaping.

LESS with GRPO. For an input query q, the policy produces G responses with rewards $\{r_1, \ldots, r_G\}$. GRPO standardizes these rewards to obtain group-relative advantages:

$$A_i = \frac{r_i - \text{mean}(r_{1:G})}{\text{std}(r_{1:G})}. \tag{5}$$

GRPO then updates the policy by maximizing a clipped likelihood-ratio objective regularized by a KL constraint toward a reference policy:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\left(\min(\alpha_i A_i,\ \tilde{\alpha}_i A_i) - \kappa_i\right)\right].$$

$$\text{where}\ \ \alpha_i = \frac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)},\ \ \tilde{\alpha}_i = \text{clip}(\alpha_i, 1-\epsilon, 1+\epsilon),$$

$$\kappa_i = \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}), \tag{6}$$

GRPO's group-wise credit assignment is well aligned with LESS, since the same group of responses used for reward normalization is also used by LESS to compute low-entropy statistics. Replacing A_i in Eq. 6 with the shaped advantage $\hat{A}_j^i$ in Eq. 4 yields our LESS-GRPO training objective.

LESS with GSPO. GSPO differs from GRPO by performing importance weighting and clipping at the sequence level. Given the same group of responses $\{o_i\}_{i=1}^G$, GSPO defines the sequence-level importance ratio and optimizes

$$J_{\text{GSPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\left(\min(\beta_i A_i,\ \tilde{\beta}_i A_i) - \kappa_i\right)\right].$$

$$\text{where}\ \ \beta_i = \left(\frac{\pi_\theta(o_i \mid x)}{\pi_{\theta_{\text{old}}}(o_i \mid x)}\right)^{\frac{1}{|o_i|}},\ \kappa_i = \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}),$$

$$\tilde{\beta}_i = \text{clip}(\beta_i, 1-\epsilon, 1+\epsilon), \tag{7}$$

Similarly, to integrate LESS with GSPO, we retain GSPO's sequence-level importance weighting but replace the scalar group advantage A_i Eq.7 with token-level shaped advantages in Eq.4.

4 Experiments

We answer the following research questions with experiments: **RQ1**:How does LESS affect performance across diverse benchmarks when applied on top of standard RLVR algorithms (GRPO and GSPO), compared with strong baselines? **RQ2**: How does LESS influence the training dynamics of LLM reasoning, compared with GRPO, in terms of accuracy growth, stability. **RQ3**: How do LESS and GRPO differ in the evolution of entropy-based reasoning structures during training? **RQ4**: Does LESS improve worst-case reasoning robustness compared with GRPO across different model sizes? **RQ5**: How sensitive is LESS to the minimum segment-length μ, and how does varying μ affect the stability and final accuracy of reinforcement-learning-based reasoning?

4.1 Experimental Setup

Datasets and evaluation metrics. Following (Shen, 2025), we train the models on the MATH dataset (Hendrycks et al., 2021), which contains 7,500 problems spanning algebra, geometry, counting, probability, number theory, and other areas. The dataset is widely adopted in LLM reasoning research due to its breadth and the step-wise reasoning it elicits, making it particularly suitable for entropy-based structure analysis.

In terms of the evaluation, we assess the trained models on a suite of standard mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME'24, and AIME'25 (LI et al., 2024). These datasets collectively cover varying difficulty levels and reasoning types, allowing us to examine whether LESS consistently improves reasoning stability. For all benchmarks except AIME, we report accuracy under greedy decoding, which is commonly used in math reasoning evaluation. For AIME'24/25, we follow prior works(Yu et al., 2025; Zheng et al., 2025; Yue et al., 2025) and compute the avg@32 accuracy by averaging predictions over 32 sampled rollouts. This protocol reduces the variance introduced by integer-answer formats and ensures fair comparison across RL-trained models.

Backbone LLM and baselines. We evaluate LESS on three Qwen2.5 family: Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, and the general-purpose Qwen2.5-7B model (Qwen et al., 2024). These models allow us to test LESS across (i) different parameter scales and (ii) models with and without domain-specific pretraining.

We compare LESS against strong RLVR systems including: **GRPO** (Shao et al., 2024), the canonical multi-sample policy-gradient method and the underlying backbone of many reasoning RL pipelines; **Forking Tokens** (Wang et al., 2025), an approach that identifies repeated reasoning fragments to adjust token-level credit assignment. **KL-Cov** (Cui et al., 2025), an entropy-based mechanism that modulates KL penalties using covariance between reward and token log-probs. These baselines represent the closest lines of work involving token-level structure, multi-sample variance reduction, and entropy-informed regularization, making them the most relevant comparisons for evaluating LESS.

Implementation details. We provide comprehensive implementation details in Appendix A.3. Our code is available at `https://anonymous.4open.science/r/LESS-2543/`.

4.2 Overall Performance

To answer **RQ1**, we conduct experiments on seven reasoning benchmarks across three Qwen2.5 backbones. Table 1 shows that LESS consistently outperforms all RL baselines on every backbone. On the math-specialized 1.5B model, LESS (GRPO) attains the highest average score, with clear gains on challenging tasks such as AIME24, AIME25, and AMC23, indicating that entropy-aware advantage shaping improves reliability over token-level clipping and forking-based updates. For the stronger 7B math model, LESS (GRPO) further raises the average to 47.7, with the largest margins on MATH500 and OlympiadBench, suggesting that the method strengthens multi-step reasoning structure rather than only local symbolic steps. On the 7B base model without math specialization, LESS (GRPO) still brings consistent gains, showing that the approach generalizes beyond math-aligned checkpoints. When instantiated with GSPO, LESS also yields improvements over vanilla GSPO across two 7B backbones, supporting its role as a generic credit-shaping module. Overall, LESS delivers the best average performance in every setting, improving both easy and hard rea-

Table 1: Overall performance on mathematical reasoning benchmarks. AIME24/25 are evaluated with avg@32; other benchmarks use @avg1. Bold numbers are the best in each column. '†' indicates the model significantly outperforms all baseline models with paired t-tests at $p < 0.05$ level.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg
Qwen2.5-Math-1.5B							
Base LLM	6.2	3.8	37.5	58.6	15.8	26.5	24.7
Forking Tokens	21.6	7.0	60.0	75.6	29.0	38.5	38.6
KL-Cov	22.0	9.2	57.5	**75.8**	27.9	38.4	38.4
GRPO	21.6	6.6	57.5	74.8	27.2	**39.6**	37.8
LESS (GRPO)	**26.2**†	**12.4**†	**62.5**†	75.2	**30.8**†	**39.6**	**41.1**†
Qwen2.5-Math-7B							
Base LLM	6.0	8.9	57.5	58.6	28.7	38.0	32.9
Forking Tokens	36.6	14.2	67.5	78.2	37.8	42.2	46.0
KL-Cov	35.2	14.1	**70.0**	78.6	38.2	43.7	46.6
GSPO	36.8	13.2	**70.0**	79.8	36.7	42.7	46.5
LESS (GSPO)	**40.0**†	13.6	67.5	79.6	37.3	44.3	47.0
GRPO	33.3	13.8	65.0	79.8	**38.6**	44.9	45.9
LESS (GRPO)	36.0	**15.6**†	**70.0**	**81.6**†	37.8	**45.7**†	**47.7**†
Qwen2.5-7B							
Base LLM	3.2	5.2	37.5	53.4	18.0	23.9	23.5
Forking Tokens	20.2	12.3	62.5	77.8	37.5	40.6	41.8
KL-Cov	18.7	10.8	60.0	77.0	**37.9**	40.8	40.9
GSPO	18.3	11.2	62.5	78.4	36.3	40.2	41.1
LESS (GSPO)	19.2	11.6	62.5	77.8	37.8	**41.8**†	41.8
GRPO	18.9	11.9	60.0	76.8	37.1	40.9	40.9
LESS (GRPO)	**20.5**†	**13.1**†	**67.5**†	**78.6**†	37.1	40.8	**42.9**†

(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B (c) Qwen2.5-7B

Figure 2: Training dynamics (accuracy over training) of GRPO and LESS across three backbones.

soning benchmarks. Unless otherwise stated, all detailed analyses in the following section use the GRPO-based instantiation of LESS.

4.3 Training Dynamics Analysis

To answer **RQ2**, we examine LESS and GRPO on how accuracy and entropy-based structure evolve throughout training. As shown in Fig. 2, across all three backbones, LESS exhibits a characteristic two-phase learning pattern. In the early stage, LESS improves slightly slower than GRPO because its advantage shaping reduces the update magnitude on low-entropy segments until the model accumulates enough evidence to distinguish productive from unproductive ones. However, as training progresses, LESS consistently surpasses GRPO and maintains a higher accuracy plateau, indicating more stable policy improvement.

To answer **RQ3**, we analyze how the overlap of correct low-entropy segments and the entropy structure of incorrect responses evolve during reinforcement learning. Fig. 3 reports training dynamics on Qwen2.5-Math-7B; similar trends on Qwen2.5-Math-1.5B and Qwen2.5-7B are presented in Appendix A.4.

The top row of Fig. 3 shows that LESS consistently yields a higher overlap ratio of correct-only low-entropy segments as training progresses. This indicates that LESS explicitly amplifies structurally productive reasoning patterns that appear repeatedly in correct trajectories. In contrast, GRPO shows a flatter trend, suggesting that it does not

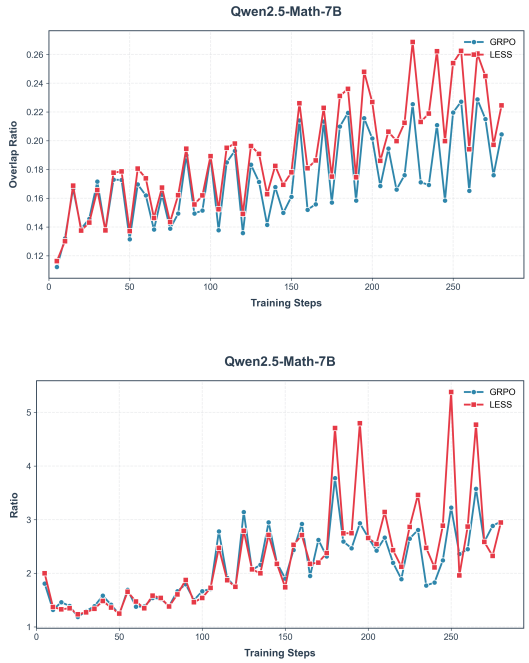


Figure 3: Training-dynamics comparison between LESS and GRPO on Qwen2.5-Math-7B. **Top**: Ratio of low-entropy segments that overlap exclusively among correct responses (higher is better). **Bottom**: Ratio between the entropy of incorrect responses and correct responses (higher indicates that incorrect answers remain exploratory). LESS consistently strengthens productive low-entropy structures while preventing premature entropy collapse in incorrect trajectories.

Table 2: Worst-case reasoning performance (worst@k) across three backbones. For each prompt, the worst-performing sample among k rollouts is selected and averaged over the dataset. LESS consistently improves worst-case accuracy across all settings.

Method	worst@32	worst@16	worst@8
Qwen2.5-Math-1.5b			
GRPO	6.8	8.0	10.8
LESS(GRPO)	**12.9**	**15.6**	**18.2**
Qwen2.5-Math-7b			
GRPO	13.4	17.4	20.6
LESS(GRPO)	**21.2**	**22.2**	**24.3**
Qwen2.5-7B			
GRPO	10.3	11.3	12.9
LESS(GRPO)	**11.1**	**12.0**	**13.4**

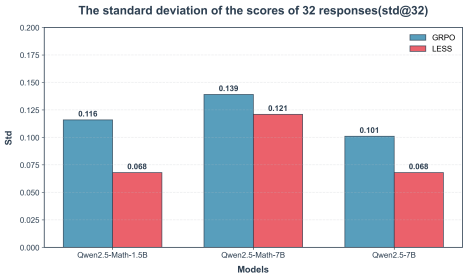


Figure 4: Standard deviation of 32 sampled responses (std@32). LESS reduces response-level variability across all backbones, indicating more stable and less volatile reasoning behavior compared with GRPO.

reliably consolidate these stable reasoning components. The clearer upward trajectory of LESS reveals that the model is progressively internalizing reusable, high-quality reasoning routines rather than relying on isolated or brittle solution paths.

The bottom row of Fig. 3 further shows that LESS maintains a higher entropy ratio between incorrect and correct responses, meaning that incorrect trajectories remain more uncertain. This separation is desirable, that is, LESS avoids prematurely stabilizing low-entropy segments that consistently lead to wrong answers, thereby reducing the risk of "locking in" systematic errors. GRPO, however, frequently collapses the entropy gap, causing incorrect responses to become low-entropy as well—an indication that harmful patterns are becoming entrenched in the policy.

These results show that LESS not only improves performance, but also progressively increases the overlap of correct low-entropy segments while keeping incorrect trajectories uncertain, creating a clear structural separation between productive and unproductive reasoning.

4.4 Robustness Under Worst-Case Sampling

To answer **RQ4** and examine the robustness of the learned policy, we employ the worst@k metric, which selects the lowest-scoring output among k sampled responses and averages this worst-case score across prompts. This metric directly measures how the model behaves in its most vulnerable failure modes.

As shown in Table. 2, across all three backbones, LESS consistently improves worst-case accuracy. For Qwen2.5-Math-1.5B and Qwen2.5-Math-7B, LESS achieves substantial gains, raising worst@32 by +6.1 and +7.8 points respectively, with positive margins maintained as k decreases. This pattern shows that LESS not only lifts average performance but also strengthens the weakest trajectories, suppressing brittle low-entropy patterns that GRPO tends to reinforce. Even on the non-math Qwen2.5-7B model, LESS produces steady improvements, indicating that its robustness effects generalize be-

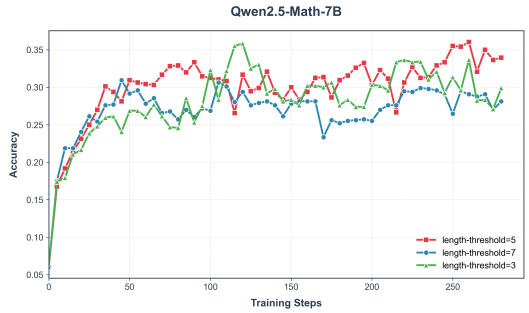


Figure 5: Effect of the low-entropy segment length threshold $\mu = \{3, 5, 7\}$ on training dynamics. We report accuracy over training steps.

yond specialized mathematical checkpoints.

The variance results shown in Fig. 4 further reinforce this finding: LESS consistently reduces the standard deviation of sampled rollouts (std@32), yielding more stable and predictable reasoning behavior. Together, the worst@k and variance metrics demonstrate that LESS raises the floor of model performance while simultaneously mitigating response-level volatility.

These results show that LESS meaningfully improves robustness by raising the floor of model performance while simultaneously reducing instability across sampled rollouts.

4.5 Impact of the Segment Length

To answer **RQ5**, we study the sensitivity of LESS to the minimum segment-length threshold μ, we vary $\mu \in \{3, 5, 7\}$ and track training dynamics on Qwen2.5-Math-7B. As shown in Figure 5, $\mu = 5$ produces the most stable and highest final accuracy across the entire training trajectory. A very small threshold ($\mu = 3$) makes the model overly sensitive to short, noisy low-entropy fragments, causing the policy to reinforce many spurious local patterns and resulting in pronounced fluctuations. Conversely, a larger threshold ($\mu = 7$) filters out too many low-entropy segments, delaying the discovery of reliable reasoning motifs and slowing convergence.

The superior performance of $\mu = 5$ suggests that effective low-entropy guidance requires a balance: segments must be long enough to encode meaningful reasoning structure, yet short enough to capture fine-grained patterns that recur across correct trajectories. This indicates that LESS benefits from moderately sized structural units and is robust to reasonable choices of μ, but extremely small or large thresholds degrade the quality of structural signals made available to the policy.

5 Related Work

Token credit assignment. Vassoyan et al. (2025) identify critical tokens in chain-of-thought solutions– decision points where the model is likely to fail, and increase exploration around these tokens by adjusting the KL penalty. Lin et al. (2024) likewise locate tokens that strongly influence incorrect outcomes and show that editing or replacing these tokens can change the final decision. Other work (Chan et al., 2024; Xie et al., 2025; Guo et al., 2025; Wei et al., 2025a) addresses the coarse-grained nature of standard RL feedback by constructing dense, token-level rewards to resolve the credit assignment problem. These methods demonstrate that tokens within a trajectory should not be treated uniformly, but they still operate on local positions and do not capture how stable patterns repeat across multiple rollouts of the same question.

Entropy-based RL signals. Wang et al. (2025) split trajectories at high-entropy tokens and update only a subset of tokens, aiming to reduce over-optimization on already confident regions. Cui et al. (2025) further modulate the KL penalty based on token-level uncertainty, encouraging updates where the model is less certain and damping updates on low-entropy tokens. In these approaches, high-entropy tokens serve as a proxy for exploration, while low-entropy regions are treated as parts of the trajectory that should be protected from change. However, they do not distinguish low-entropy patterns that are consistently correct from those that encode repeated mistakes.

6 Conclusion

This paper presents a new perspective on training reasoning LLMs: reasoning should be guided at the level of low-entropy segments. Building on this insight, we propose LESS, a plug-and-play advantage-shaping framework that selectively amplifies reliable low-entropy reasoning segments and suppresses error-prone ones. Instantiated with GRPO and GSPO, LESS improves accuracy, stability, and robustness across multiple backbones and benchmarks. The framework consistently strengthens correct reasoning routines, preserves exploration on incorrect trajectories, and raises the worst-case performance of sampled rollouts. These results show that low-entropy structural signals offer a principled and effective handle for guiding RL training of reasoning models.

7 Limitations

We believe there is still room for improvement in our work. Our preliminary analysis and all main experiments are conducted on the Qwen2.5 family. We do not test LESS on other backbones such as Llama, so it is unclear whether the same entropy patterns and gains will hold more broadly. In addition, following Wang et al. (2025), we fix the entropy quantile and length threshold to extract low-entropy segments. We do not yet study how different quantiles, adaptive thresholds, or alternative segmentation rules would affect the learned segments and the final performance.

References

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. *arXiv e-prints*, arXiv:2508.06471.

Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Dense Reward for Free in Reinforcement Learning from Human Feedback. *arXiv e-prints*, arXiv:2402.00782.

Zhipeng Chen, Kun Zhou, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2024. Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment. *arXiv e-prints*, arXiv:2406.12606.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with Exploration: An Entropy Perspective on Reinforcement Learning for LLMs. *arXiv e-prints*, arXiv:2506.14758.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv e-prints*, arXiv:2505.22617.

Jeff Da, Clinton Wang, Xiang Deng, Yuntao Ma, Nikhil Barhate, and Sean Hendryx. 2025. Agent-RLVR: Training Software Engineering Agents via Guidance and Environment Rewards. *arXiv e-prints*, arXiv:2506.11425.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv e-prints*, arXiv:2501.12948.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. 2025. Segment Policy Optimization: Effective Segment-Level Credit Assignment in RL for Large Language Models. *arXiv e-prints*, arXiv:2505.23564.

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. 2024. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. *arXiv e-prints*, arXiv:2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv e-prints*, arXiv:2103.03874.

Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. 2025. Reason Like a Radiologist: Chain-of-Thought and Reinforcement Learning for Verifiable Report Generation. *arXiv e-prints*, arXiv:2504.18453.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving Quantitative Reasoning Problems with Language Models. *arXiv e-prints*, arXiv:2206.14858.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. `https://huggingface.co/AI-MO/NuminaMath-CoT`.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. 2024. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM's Reasoning Capability. *arXiv e-prints*, arXiv:2411.19943.

Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. 2025. GHPO: Adaptive Guidance for Stable and Efficient LLM Reinforcement Learning. *arXiv e-prints*, arXiv:2507.10628.

M2 Team, Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, Xiangrong Zeng, Yijie Zhou, Chenzheng Zhu, Da Pan, Fei Deng, Guangwei Ai, Guosheng Dong, Hongda Zhang, and 15 others. 2025. Baichuan-M2: Scaling Medical Capability with Large Verifier System. *arXiv e-prints*, arXiv:2509.02208.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 Technical Report. *arXiv e-prints*, arXiv:2412.15115.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. MedGemma Technical Report. *arXiv e-prints*, arXiv:2507.05201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv e-prints*, arXiv:2402.03300.

Han Shen. 2025. On Entropy Control in LLM-RL Algorithms. *arXiv e-prints*, arXiv:2509.03493.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv e-prints*, arXiv:2409.19256.

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. 2025. Ignore the KL penalty! boosting exploration on critical tokens to enhance RL fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6108–6118, Albuquerque, New Mexico. Association for Computational Linguistics.

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. 2025. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *arXiv e-prints*, arXiv:2506.01939.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. 2025a. Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Reward Design. *arXiv e-prints*, arXiv:2505.11821.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. 2025b. SWE-RL: Advancing LLM Reasoning via Reinforcement Learning on Open Software Evolution. *arXiv e-prints*, arXiv:2502.18449.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. 2025. CAPO: Towards Enhancing LLM Reasoning through Generative Credit Assignment. *arXiv e-prints*, arXiv:2508.02298.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv e-prints*, arXiv:2409.12122.

Kun Yang, Zikang chen, Yanmeng Wang, and Zhigen Li. 2025. SSPO: Subsentence-level Policy Optimization. *arXiv e-prints*, arXiv:2511.04256.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv e-prints*, arXiv:2503.14476.

Yu Yue, Yufeng Yuan, Qiying 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, and 8 others. 2025. VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. *arXiv e-prints*, arXiv:2504.05118.

Kepu Zhang, Guofu Xie, Weijie Yu, Mingyue Xu, Xu Tang, Yaxin Li, and Jun Xu. 2025a. Legal mathematical reasoning with LLMs: Procedural alignment through two-stage reinforcement learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1586–1598, Suzhou, China. Association for Computational Linguistics.

Kepu Zhang, Weijie Yu, Zhongxiang Sun, and Jun Xu. 2025b. Syler: A framework for explicit syllogistic legal reasoning in large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 4117–4127, New York, NY, USA. Association for Computing Machinery.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. 2025. EDGE-GRPO: Entropy-Driven GRPO with Guided Error Correction for Advantage Diversity. *arXiv e-prints*, arXiv:2507.21848.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. 2025. First Return, Entropy-Eliciting Explore. *arXiv e-prints*, arXiv:2507.07017.

A Appendix

A.1 Algorithm and Time Complexity Analysis

Algorithm 1 LESS: Low-Entropy Segment Shaping

Input:
Group of responses $\mathcal{G} = \{O_1, \ldots, O_G\}$, token advantages $\mathcal{A} = \{A_j^i \mid t_j \in O_i, i = 1, \ldots, G\}$, token entropies $H = \{\mathcal{H}_j^i \mid t_j \in O_i, i = 1, \ldots, G\}$, correctness labels $\{correct_1, \ldots, correct_G\}$, entropy quantile h, minimum segment length μ.
Output: Shaped advantages $\mathcal{A}' = \{\hat{A}_j^i \mid t_j \in O_i, i = 1, \ldots, G\}$
$N_r \leftarrow \sum_i \mathbb{I}[correct_i = 1]$; $N_w \leftarrow \sum_i \mathbb{I}[correct_i = 0]$
for each response O_i **do**
  Compute entropy threshold τ_i from $\{\mathcal{H}_j^i\}_{t_j \in O_i}$ using quantile h
  Segment O_i into $\mathcal{S}_i^{\text{high}}$, $\mathcal{S}_i^{\text{frag}}$, $\mathcal{S}_i^{\text{seg}}$ using τ_i and μ (Eq. (2))
end for
$\Sigma \leftarrow \varnothing$ {set of unique low-entropy segments}

for each response O_i **do**
  for all $\sigma \in \mathcal{S}_i^{\text{seg}}$ **do**
    if no $\sigma' \in \Sigma$ is a contiguous segments of σ **then**
      $\Sigma \leftarrow \Sigma \cup \{\sigma\}$
      Remove from Σ any σ' that is strictly contained in σ
    end if
  end for
end for
for all $\sigma \in \Sigma$ **do**
  Compute $n_r(\sigma)$ and $n_w(\sigma)$ according to Eq. (3)
end for
for each response O_i **do**
  for each token $t_j \in O_i$ **do**
    Set $\hat{A}_j^i$ according to Eq. (4)
  end for
end for
return $\mathcal{A}' = \{\hat{A}_j^i\}$

The overall LESS procedure is summarized in Algorithm 1. Given a group of responses and their token-level entropies, we first compute an entropy threshold for each response and segment it into high-entropy tokens, short low-entropy fragments, and longer low-entropy segments (Eq. (2)). We then build a set Σ of non-redundant low-entropy segments across the group by keeping only segments that are not strictly contained in longer ones. For each segment $\sigma \in \Sigma$, we count how many correct and incorrect responses it appears in (Eq. (3)), and finally assign a shaped advantage to every token based on its entropy category and the statistics of the segment it belongs to (Eq. (4)). The resulting token-level advantages $\mathcal{A}'$ can be plugged into any group-based RLVR update.

In terms of complexity, when the batch size is B, the group size is G, and the maximum response length is L, the segmentation and shaping operations visit each token a constant number of times, giving a practical time complexity of $O(BGL)$. Under our main setting (batch size 512, group size 8, average response length about 800), LESS adds roughly 60 seconds of overhead in our implementation, which is small compared to the overall RL training time.

Table 3: Prompt template used for all experiments. `{question}` is replaced by the problem description.

Prompt Template

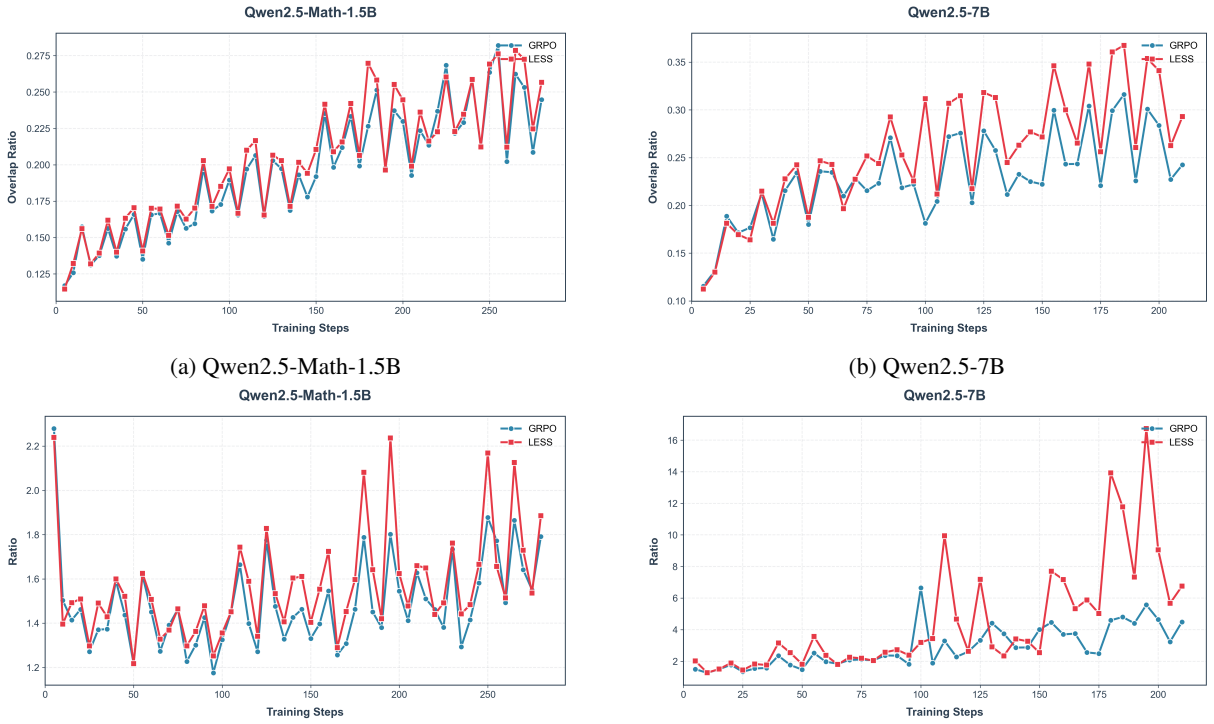
```
<|im start|>system
Please reason step by step, and put your final
answer within \boxed{}.
<|im end|>
<|im start|>user
{question}
<|im end|>
<|im start|>assistant
```

A.2 Prompt Template

We use a unified prompt template for all training and evaluation experiments, adapted from the official Qwen-Math template (Yang et al., 2024). The concrete format is shown in Table 3.

A.3 Implementation Details

We conduct experiments using the VeRL (Sheng et al., 2024) framework for reinforcement learning with LLMs on 8 NVIDIA A100-40G GPUs. The training setup includes a batch size of 512, a learning rate of 1×10^{-6}, and a clip range between 0.2 and 0.28. Each response sequence is up to 3k tokens in length. The mini-batch size is set to 32. The temperature is 1.0 for training and 0.1 for evaluation. Following prior RLVR work (Liu et al., 2025; Yang et al., 2025), we perform 8 rollouts per promp and do not use entropy regularization or



(a) Qwen2.5-Math-1.5B (b) Qwen2.5-7B

(c) Qwen2.5-Math-1.5B (d) Qwen2.5-7B

Figure 6: Training-dynamics comparison between LESS and GRPO across two model sizes. **Top**: Ratio of low-entropy segments that overlap exclusively among correct responses (higher is better). **Bottom**: Ratio between the entropy of incorrect responses and correct responses (higher indicates that incorrect answers remain exploratory).

KL penalties during training (KL coefficient = 0, entropy loss = 0), allowing us to isolate the effect of LESS from other types of regularization. This choice also reveals whether LESS alone can stabilize reasoning trajectories without relying on heavy KL anchoring. We use a unified prompt format for all experiments, and the exact template is provided in Appendix A.2.

A.4 Additional Training Dynamics Analysis

Fig. 6 presents the same training-dynamics analysis for Qwen2.5-Math-1.5B and Qwen2.5-7B. Across both backbones, LESS consistently increases the overlap of correct-only low-entropy segments during training and maintains a higher entropy ratio between incorrect and correct responses. These trends mirror the behavior observed on Qwen2.5-Math-7B, indicating that the structural effects of LESS are stable across model scales and are not specific to a single backbone.

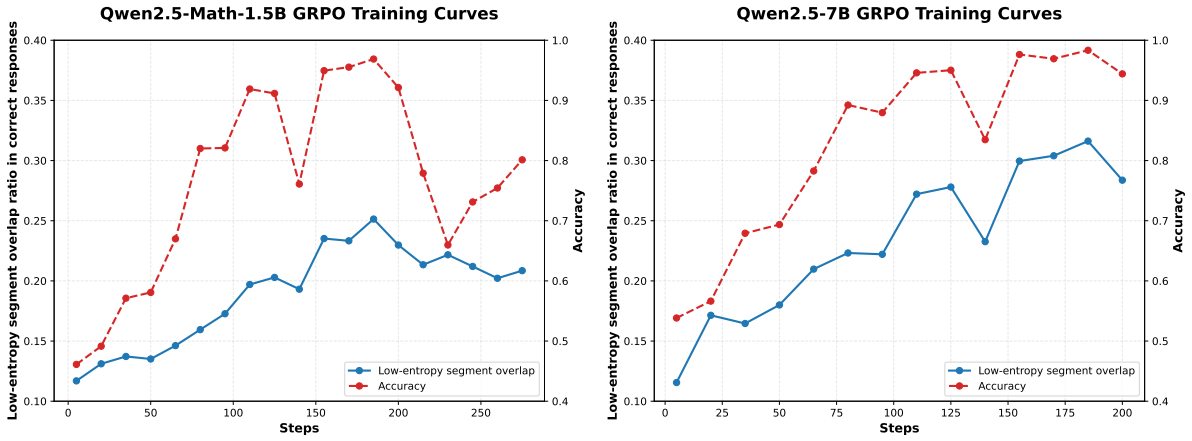
A.5 Additional Overlap–Accuracy Curves

To check whether the correlation between low-entropy overlap and accuracy holds beyond Qwen2.5-Math-7B, we repeat the analysis in § 2 on Qwen2.5-Math-1.5B and Qwen2.5-7B. Figure 7 shows that, under GRPO training, the overlap ratio of low-entropy segments in correct responses grows together with accuracy for both backbones. Early in training, both curves are low and noisy; as learning proceeds, the overlap becomes higher and smoother while accuracy also rises. These results support our claim that low-entropy segment overlap tracks the formation of stable reasoning routines across different model sizes and pretraining setups.

A.6 Additional Accuracy–Overlap Correlations

In § 2, we report that, for Qwen2.5-Math-7B, benchmark accuracy is strongly correlated with the overlap of low-entropy segments across correct responses. To test the robustness of this phenomenon, we repeat the correlation analysis on four additional backbones: Qwen2.5-7B, DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B, and Qwen2.5-Math-1.5B-Oat-Zero.

Figure 8 summarizes the results. For each backbone, we compute four overlap ratios at the benchmark level: (i) overlap among all responses, (ii) overlap among correct responses only, (iii) overlap among segments shared by correct and incorrect responses, and (iv) overlap among incorrect



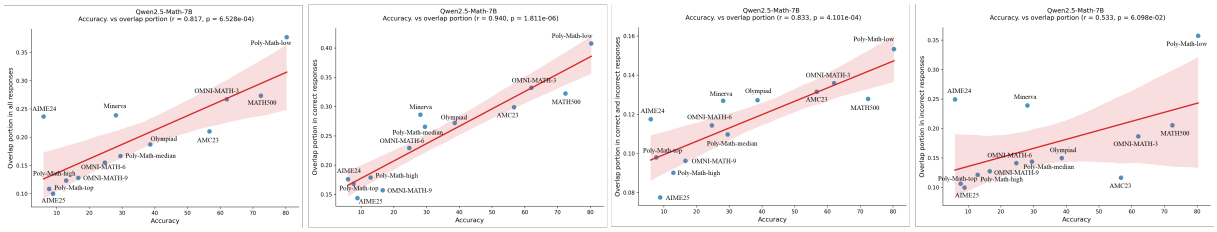
(a) Qwen2.5-1.5B-Math

(b) Qwen2.5-7B

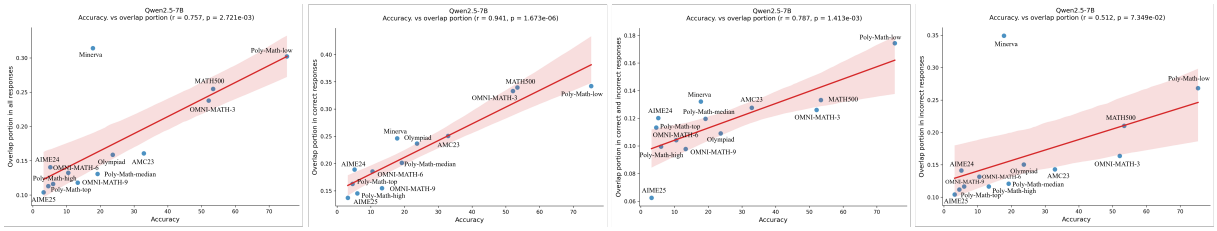
Figure 7: GRPO training curves on two additional backbones. We plot the overlap ratio of low-entropy segments in correct responses (left y-axis, blue) and accuracy (right y-axis, red) over training steps for (a) Qwen2.5-Math-1.5B and (b) Qwen2.5-7B. In both models, low-entropy overlap and accuracy increase in tandem, echoing the trend observed for Qwen2.5-Math-7B in the main text.

responses only. We then correlate each ratio with benchmark accuracy.

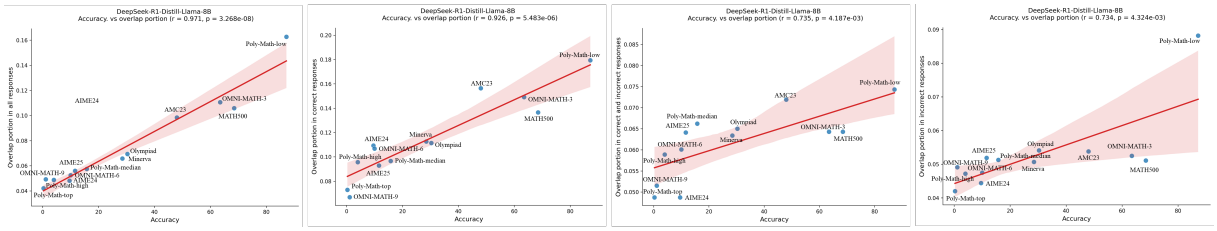
Across all five backbones, we observe a consistent pattern: overlap among correct responses shows the strongest positive correlation with accuracy, overlap among all responses and shared segments yields weaker but still positive correlations, and overlap among incorrect-only segments is weakly correlated or even negatively correlated. These additional results support our claim that stable low-entropy structure in correct trajectories is a reliable indicator of reasoning quality, while overlap driven by incorrect trajectories does not translate into better performance.



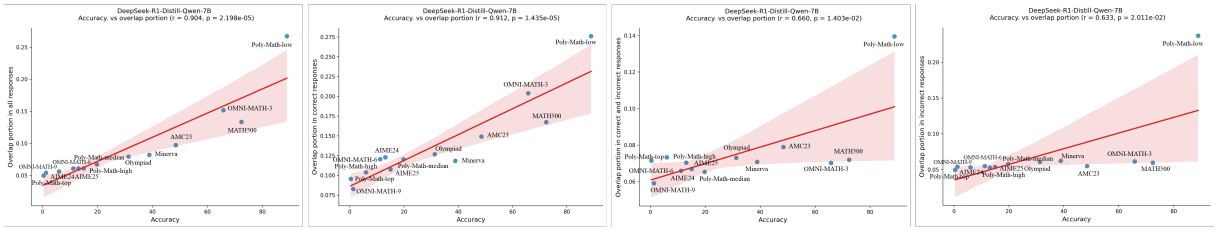
(a) Qwen2.5-Math-7B



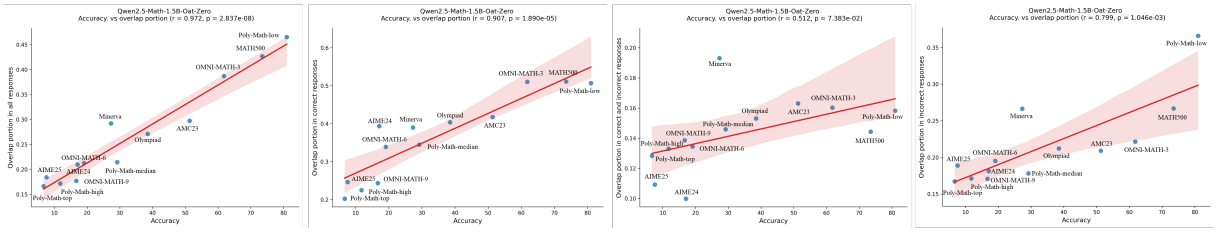
(b) Qwen2.5-7B



(c) DeepSeek-R1-Distill-Llama-8B



(d) DeepSeek-R1-Distill-Qwen-7B



(e) Qwen2.5-Math-1.5B-Oat-Zero

Figure 8: Additional correlations between accuracy and low-entropy segment overlap across backbones. Panels (a)–(e) report, for Qwen2.5-Math-7B, Qwen2.5-7B, DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B, and Qwen2.5-Math-1.5B-Oat-Zero, the Pearson correlations between benchmark accuracy and four overlap ratios: all responses, correct-only responses, segments shared by correct and incorrect responses, and incorrect-only responses. Each point is a benchmark; the red line and shaded area show the fitted regression and its confidence band. Across models, accuracy is most strongly aligned with overlap among correct responses.