Efficient Reinforcement Learning for Large Language Models with Intrinsic Exploration

Yan Sun[1,2*] **Jia Guo**[2†] **Stanley Kok**[1] **Zihao Wang**[2] **Zujie Wen**[2] **Zhiqiang Zhang**[2]
[1]National University of Singapore [2]Ant Group
{yansun, skok}@comp.nus.edu.sg
{jia.g, xiaohao.wzhg, zujie.wzj, lingyao.zzq}@antgroup.com

Abstract

Reinforcement learning with verifiable rewards (RLVR) has improved the reasoning ability of large language models, yet training remains costly because many rollouts contribute little to optimization, considering the amount of computation required. This study investigates how simply leveraging intrinsic data properties, almost free benefit during training, can improve data efficiency for RLVR. We propose PREPO with two complementary components. First, we adopt prompt perplexity as an indicator of model adaptability in learning, enabling the model to progress from well-understood contexts to more challenging ones. Second, we amplify the discrepancy among the rollouts by differentiating their relative entropy, and prioritize sequences that exhibit a higher degree of exploration. Together, these mechanisms reduce rollout demand while preserving competitive performance. On the Qwen and Llama models, PREPO achieves effective results on mathematical reasoning benchmarks with up to 3 times fewer rollouts than the baselines. Beyond empirical gains, we provide theoretical and in-depth analyses explaining the underlying rationale of our method to improve the data efficiency of RLVR.

1 Introduction

Reinforcement learning (RL) has become central to improving the reasoning capabilities of large language models (LLMs) by optimizing their self-generated rollouts [Guo et al., 2025, Team et al., 2025]. Recent advances in reinforcement learning with verifiable reward (RLVR) demonstrate that it is a simple yet effective method for scaling reasoning performance [Shao et al., 2024, Yu et al., 2025]. However, RLVR still incurs substantial computational overhead, as rollout generation remains the primary training bottleneck [Zhong et al., 2024].

A key source of inefficiency is that not all samples contribute equally to training. On the *prompt side*, some queries are too trivial or too difficult to contribute meaningful gradient [Yu et al., 2025]. Prompt difficulty is often estimated through pass rates [Zhang et al., 2025] or manually defined criteria [Chen et al., 2025a, Parashar et al., 2025], but these approaches are expensive and may not reflect the model's own perception. On the *rollout side*, responses differ in confidence regardless of correctness (see Fig. 7). Low-entropy (confident) responses produce small gradients, whereas highly-entropy (uncertain) responses produce large ones, implying alternative reasoning paths that support exploration (see Appendix A). Thus, we may leverage the intrinsic property as an inductive bias for more efficient training.

A natural way to improve efficiency is through data selection, i.e., pruning uninformative prompts or rollouts while preserving those that drive learning. There are emerging approaches based on

[*]Work done during internship at Ant Group
[†]Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

parameterized modeling [Qu et al., 2025], replay buffers [Liu et al., 2025], or selective rollout execution [Zheng et al., 2025]. This motivates our research question from a new perspective:

Can the intrinsic properties of prompts and rollouts improve the efficiency of RLVR?

In this study, we propose Perplexity-Schedule with Relative-Entropy Policy Optimization (PREPO)[3], which combines a perplexity-based schedule with sequence-level entropy weighting to realize *intrinsic exploration*. Specifically, PREPO traces perplexity *before* rollout generation to pruning the prompts, and applies entropy weighting *after* rollout generation to emphasize uncertain responses. Across Qwen and Llama models, PREPO surpasses existing data-pruning baselines and remains competitive with the full-data setting, while reducing rollout usage by more than 40% (see Fig. 1 for Qwen2.5-Math-7B). These results show that RLVR can be made substantially more efficient by leveraging the intrinsic properties of prompt and rollout data.

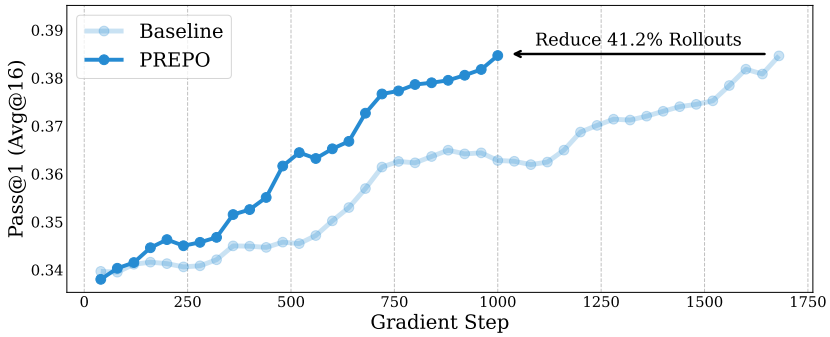


Figure 1: Comparison of PREPO and GRPO with random 20% selection on Qwen2.5-Math-7B, averaged across AIME24, AIME25, MATH-500, and Olympiad Bench.

2 Preliminary Analysis

2.1 Prompts with Lower-PPL Tend to Yield Higher Passrate

We begin by examining the relationship between prompt perplexity (PPL) and task difficulty using the DAPO-Math-17K dataset[Yu et al., 2025]. For both Qwen and Llama models, Figure 2 shows a clear negative correlation between PPL and passrate@16, where passrate@16 measures the fraction of prompts solved by at least one of 16 generations. Prompts with lower PPL generally yield higher success rates. Table 1 correlation is statistically significant across models, suggesting that PPL can serve as a lightweight signal to identify more informative prompts for training.

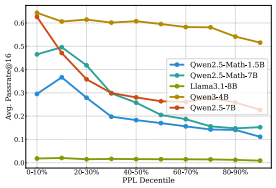


Figure 2: Prompt PPL versus average passrate@16.

Table 1: Correlation between prompt PPL and *passrate@16*. (*** $p < 0.001$, ** $p < 0.05$)

	Qwen2.5-7B	Qwen2.5-32B	Qwen2.5-7B	Qwen2.5-1.5B	Qwen3-4B	LLama3.1-8B
Spearman	-0.233^{***}	-0.207^{***}	-0.183^{**}	-0.186^{***}	-0.169^{***}	-0.199^{***}

2.2 Training Dynamics of Low-PPL and High-PPL Prompts

We further compare the training dynamics between the LOW-PPL group (i.e., data with the lowest 20% prompt perplexity) and the HIGH-PPL group (i.e., data with the highest 20% prompt perplexity) from the entire DAPO-Math-17K dataset. As illustrated in Figure 3, the two groups show distinct behaviors in multiple metrics, where LOW-PPL prompts provide stronger learning signals at the early steps, whereas HIGH-PPL prompts retain exploratory benefits that can improve sample efficiency in later steps. Please refer to Appendix B for a detailed analysis of other Qwen and Llama models.

[3]Github Repository: https://github.com/yan-sun-x/PREPO

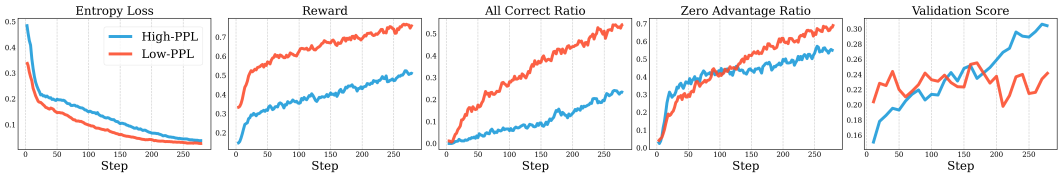


Figure 3: Training dynamics of LOW-PPL vs. HIGH-PPL prompts on Qwen2.5-Math-7B. *(a)* HIGH-PPL *prompts have higher entropy. (b)* LOW-PPL *prompts have more reward gains. (c)* LOW-PPL *prompts reach higher all-correct ratios faster. (d)* LOW-PPL *prompts show higher zero-advantage ratios in the later stage. (e)* HIGH-PPL *prompts eventually outperform* LOW-PPL *prompts on AIME24 [Art of Problem Solving, 2024].*

As a baseline, we randomly sample 20% of the data. As shown in Figure 4, the HIGH-PPL and LOW-PPL groups show distinct behaviors compared to random selection, indicating that PPL-based grouping offers a useful data pruning strategy.

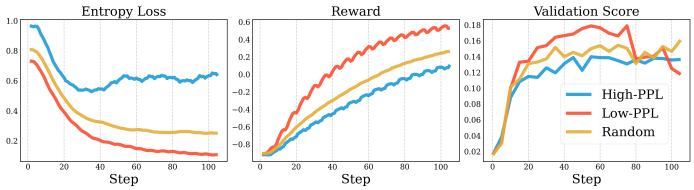


Figure 4: Comparison among LOW-PPL, HIGH-PPL, and Random Subsets. *Random lies between the two, showing that PPL-based grouping provides a meaningful pruning signal.*

3 PREPO: PPL-Schedule Relative-Entropy Policy Optimization

3.1 General Online Batch Selection

Let $\mathcal{B} = \{x_i\}_{i=1}^{|\mathcal{B}|}$ denote the candidate batch at a training step. The goal of online batch selection is to design a mapping

$$\Phi : [0, 1] \to 2^{\mathcal{B}}, \quad \rho \mapsto \mathcal{I}_\rho, \tag{1}$$

where $\rho \in [0, 1]$ denotes the normalized training progress, and $\mathcal{I}_\rho \subseteq \mathcal{B}$. The mapping Φ is required to (i) explicitly depend on ρ, so that the distribution of selected samples evolves with training; (ii) the sub-batch size is fixed during training, i.e., $\forall \rho,\ |\mathcal{I}_\rho| = K$.

3.2 PPL-Schedule Online Batch Selection

For a prompt $x_i = (x_{i,1}, \ldots, x_{i,T})$, the perplexity at progress ρ is

$$P_i(\rho) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log \pi_\rho(x_{i,t} \mid x_i, x_{i,<t})\right), \tag{2}$$

where π_ρ is the model distribution at progress ρ. Since π_ρ evolves with training, $P_i(\rho)$ reflects a dynamic difficulty score of the problem. We define the *PPL-schedule* sub-batch as

$$\mathcal{I}_\rho = \{\, \sigma(j) : l(\rho) \le j \le l(\rho) + K - 1 \,\}, \tag{3}$$

where σ is the permutation that sorts $\mathcal{B}$ by ascending $P_i(\rho)$. The starting index $l(\rho)$ is given by a linear schedule

$$l(\rho) = \lfloor \rho \cdot (|\mathcal{B}| - K) \rfloor, \tag{4}$$

so that $\mathcal{I}_\rho$ shifts smoothly from prompts with lower PPL to those with higher PPL. While linear scheduling is the simplest case, nonlinear pacing (e.g., quadratic or exponential) can also be used.

3.3 Relative-Entropy Weighting

Empirically, we find that training on low-PPL prompts accelerates reward improvement but also leads to a rapid collapse of entropy, thereby reducing exploration. To mitigate this effect during the PPL-schedule, we introduce a sequence-level relative-entropy weighting scheme that adaptively emphasizes uncertain rollouts.

The token-level entropy of a rollout is defined as $H_t = -\sum_{v\in\mathcal{V}} \pi_\theta(v \mid o_{<t}, x) \log \pi_\theta(v \mid o_{<t}, x)$, where $\mathcal{V}$ is the vocabulary. For rollout i, the sequence-level entropy is the average across its tokens

$$\bar{H}_i = \bar{H}(o_i \mid x) = \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} H_t. \tag{5}$$

The batch-average entropy over B rollouts is

$$\bar{H} = \frac{1}{B} \sum_{k=1}^{B} \bar{H}_k. \tag{6}$$

The relative weight assigned to rollout i is then given by

$$w_i = \frac{\bar{H}_i}{\bar{H}}. \tag{7}$$

This formulation ensures that weights are scale-invariant, where rollout's contribution is determined only by how its entropy compares to the batch mean. Under mild assumptions (see Appendix E), the total weight $\frac{1}{B}\sum_i w_i$ remains unbiased with respect to the actual batch size B.

Intuitively, this design enables the model to *seek uncertainty within certainty* during the PPL-schedule. While the prompts with lower PPL at early stages typically produce confident (i.e., low-entropy) responses, the weighting mechanism selectively amplifies relatively uncertain (i.e., high-entropy) rollouts, thereby preserving exploratory capacity throughout training.

3.4 Objective Function

The PREPO objective integrates PPL-schedule filtering with relative-entropy weighting as below.

$$\mathcal{J}_{\text{PREPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{I}_\rho,\ \{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} w_i \cdot \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(s_{i,t}(\theta)\, \hat{A}_{i,t},\ \text{clip}\big(s_{i,t}(\theta),\, 1-\epsilon_{\text{low}},\, 1+\epsilon_{\text{high}}\big)\, \hat{A}_{i,t} \Big) \right] \tag{8}$$

where $\mathcal{I}_\rho$ is the PPL-schedule-filtered batch of prompts at training progress ρ, w_i encodes the relative entropy of rollout i at the current micro-batch, $s_{i,t}(\theta)$ is the token-level importance ratio,$s_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|x,o_{i,<t})}{\pi_{\text{old}}(o_{i,t}|x,o_{i,<t})}$, and $\hat{A}_{i,t}$ is the group-based advantage estimate, $\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}$.

4 Related Work

Data efficiency in RLVR. There have been increasing attention to improve data efficiency for RLVR. Several works focus on prompt-side selection. For example, online difficulty filtering Bae et al. [2025], predictive prompt selection Qu et al. [2025], and curriculum-based methods Zhang et al. [2025], Chen et al. [2025a] aim to allocate resources toward more learnable prompts. Other approaches explore importance-based or gradient-informed selection, such as gradient-alignment methods Kamalloo et al. [2025], influence estimation Chen et al. [2025b], or policy-advantage based prioritization Wang and Guofeng [2025]. Parallel efforts reduce redundancy in rollouts, including down-sampling strategies Li et al. [2025] and efficient replay buffer designs Liu et al. [2025]. These methods highlight the importance of identifying which data truly drives learning.

Entropy Mechanism in RLVR. Entropy has long been studied in reinforcement learning, where entropy-regularized objectives are used to promote exploration in control settings. Recent work extends this perspective to RLVR for reasoning LLMs. Cui et al. [2025] show that rapid entropy collapse is a key failure mode, and propose covariance-based updates to slow down decay. Wang et al. [2025] further find that high-entropy "forking tokens" constitute a small minority of steps yet drive most reasoning improvements, suggesting that entropy signals informativeness at the token level.

5 Experiments

Models and Datasets. We benchmark our method against random selection and GRESO [Zheng et al., 2025] across multiple model scales. Specifically, we consider Qwen2.5-7B [Team, 2024], Qwen2.5-Math-1.5B, Qwen2.5-Math-7B [Yang et al., 2024], Qwen3-4B (non-thinking) [Yang et al., 2025], and Llama3.1-8B [Dubey et al., 2024]. For training data, we include DAPO [Yu et al., 2025] and MATH500 datasets [Lightman et al., 2023a].

Training and Evaluation. All models are trained using the `verl` [Sheng et al., 2025], with `vLLM` [Kwon et al., 2023] employed for rollout generation to ensure efficient inference. For the Qwen models, evaluation is conducted on a comprehensive suite of benchmarks, including AIME25 [Art of Problem Solving, 2025], AIME24 [Art of Problem Solving, 2024], MATH500 [Lightman et al., 2023b], and OlympiadBench [He et al., 2024], which cover a diverse range of mathematical reasoning challenges. The Llama model is evaluated on MATH500 [Lightman et al., 2023b] and GSM8K [Cobbe et al., 2021]. We evaluate all models using *pass@1 (avg16)*, i.e., the accuracy of the top-1 response averaged over 16 generations, with temperature 1.

Experiment Configuration. We set the clipping thresholds with $\epsilon_{\text{low}} = 0.2$ by default and a larger $\epsilon_{\text{high}} = 0.28$ for the upper bound. For the Qwen2.5-Math models, we use a maximum context length of 4096 tokens, matching their supported limit. For Qwen3-4B and Llama3.1-8B, we set the context length to 32,768 tokens. Rollouts are generated with temperature $T = 1$ using vLLM, producing 8 responses per prompt. The global batch size is 1280, with a reduced batch size fixed at 256 and a mini-batch size of 64. For both PREPO and the baseline, we adopt an online selection ratio of $K/\mathcal{B} = 20\%$ at each training step. For GRESO, we set the targeted zero-variance percentage as 50%. The actor model is optimized with AdamW using a constant learning rate of 1×10^{-6}, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 0.01. Following Yu et al. [2025], we omit the KL-divergence regularization term. Training is applied only to the actor parameters and parallelized with Fully Sharded Data Parallel. All experiments are conducted on 32 GPUs.

5.1 Main Results

Table 2: Performance comparison (%) on Qwen. *Best results are highlighted in bold or underlined.*

Method	AIME25	AIME24	MATH	Olympiad	Avg ↑	# Rollouts ↓
Qwen2.5-7B	1.46	5.31	63.35	33.09	25.80	–
+ Random	6.98	<u>16.41</u>	75.70	38.47	34.39	716K
+ GRESO	9.22	10.83	<u>76.65</u>	<u>42.07</u>	34.59	680K
+ PREPO (Ours)	<u>10.21</u>	16.09	76.30	39.85	**35.61**	**304K**
Qwen2.5-Math-1.5B	1.88	2.97	27.85	27.41	15.03	–
+ Random	<u>20.00</u>	16.67	76.25	30.50	35.86	3.0M
+ GRESO	15.38	<u>20.00</u>	<u>76.65</u>	24.17	34.16	2.5M
+ PREPO (Ours)	<u>20.00</u>	16.67	76.25	<u>32.00</u>	**36.23**	**1.1M**
Qwen2.5-Math-7B	1.56	3.70	20.70	39.56	16.38	–
+ Random	10.00	<u>26.67</u>	77.80	<u>43.26</u>	39.45	905K
+ GRESO	18.33	25.83	77.80	26.83	37.46	654K
+ PREPO (Ours)	<u>12.81</u>	26.15	<u>77.85</u>	41.58	**39.59**	**540K**
Qwen3-4B	30.00	53.33	94.10	52.67	57.53	–
+ Random	60.00	70.00	96.00	59.33	71.33	553K
+ GRESO	56.67	69.17	96.40	57.33	69.89	472K
+ PREPO (Ours)	<u>66.67</u>	<u>80.00</u>	<u>96.60</u>	<u>60.67</u>	**75.99**	**348K**

PREPO achieves up to $3\times$ (66.7%) rollout reduction while matching or surpassing baseline performance across all models. As shown in Table 2 and 3, PREPO reduces rollout amount by $3\times$ (63.3%) on Qwen2.5-Math-1.5B, $1.7\times$ (40.3%) on Qwen2.5-Math-7B, $1.6\times$ (37.1%) on Qwen3-4B (non-thinknig), $2.4\times$ (57.5%) on Qwen2.5-7B, and $2\times$ (48.9%) on Llama3.1-8B. These results demonstrate that PREPO consistently improves data efficiency, often reducing rollout demand by two to three times, without sacrificing, and in many cases even improving, benchmark performance. More analyses, including ablations, are provided in Appendix D.

Table 3: Performance comparison (%) on Llama. *Best results are highlighted in bold or underlined.*

Method	**GSM8K**	**MATH**	**Avg ↑**	**# Rollouts ↓**
Llama3.1-8B	9.53	6.05	7.79	–
+ Random	46.63	14.60	30.61	266K
+ GRESO	41.77	16.80	29.29	273K
+ PREPO (Ours)	<u>51.10</u>	<u>21.81</u>	**36.55**	**115K**

Due to space constraints, additional discussion and analysis of PREPO are provided in Appendix 6 and Appendix E.

5.2 Ablation Study

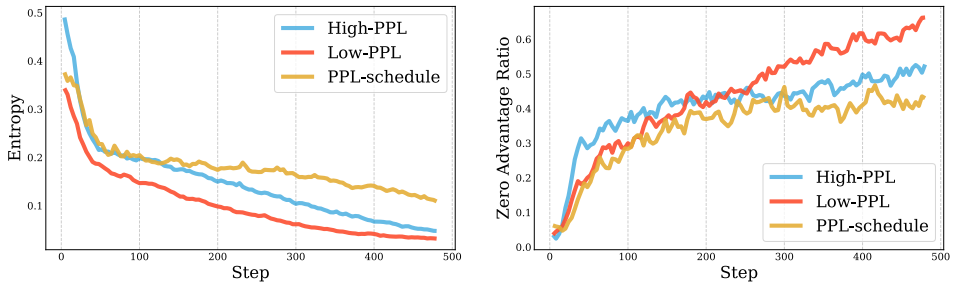
We conduct an ablation study to isolate the effect of relative-entropy weighting on top of the PPL-schedule. We report the performance of PPL-schedule at the same number of rollouts as PREPO. As shown in Table 4, PREPO consistently outperforms the PPL-schedule across most benchmarks, indicating that entropy-based rollout weighting provides additional gains beyond prompt-side scheduling.

Table 4: Performance comparison (%) between PPL-chedule and PREPO. *Best results are highlighted in bold or underlined.*

Model	Method	AIME25	AIME24	MATH	Olympiad Bench	Avg ↑
Qwen2.5-Math-7B	PPL-schedule	10.00	23.33	74.60	39.21	36.79
	+ Relative-entropy (PREPO)	<u>12.81</u>	<u>26.15</u>	<u>77.80</u>	<u>41.58</u>	**39.59**
Qwen2.5-7B	PPL-schedule	6.98	<u>16.41</u>	75.70	38.47	34.39
	+ Relative-entropy (PREPO)	<u>10.20</u>	16.09	<u>76.30</u>	<u>39.85</u>	**35.61**

6 Discussion

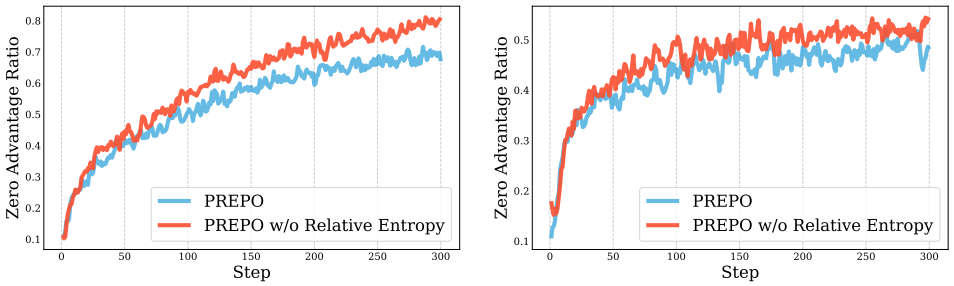
What Does the PPL-Schedule Contribute to Training? PPL-schedule helps maintain an adequate degree of exploration throughout learning. As illustrated in Figure 5, we compare three training configurations for Qwen2.5-Math-7B: (1) training exclusively with high-PPL prompts, (2) exclusively with low-PPL prompts, and (3) a PPL-schedule that gradually transitions from low- to high-PPL prompts. The PPL-schedule yields a more balanced optimization trajectory than either static regime. In terms of entropy loss, PPL-schedule shows a slower decline, avoiding the rapid entropy collapse that typically occurs under low-PPL-only training. Regarding the zero-advantage ratio, the PPL-schedule consistently achieves lower values, meaning that a larger proportion of rollouts provide non-trivial gradient contributions.



(a) Entropy (b) Zero Advantage Ratio

Figure 5: Comparison of training dynamics between PPL-schedule and static PPL selection (Low- and High-PPL groups)

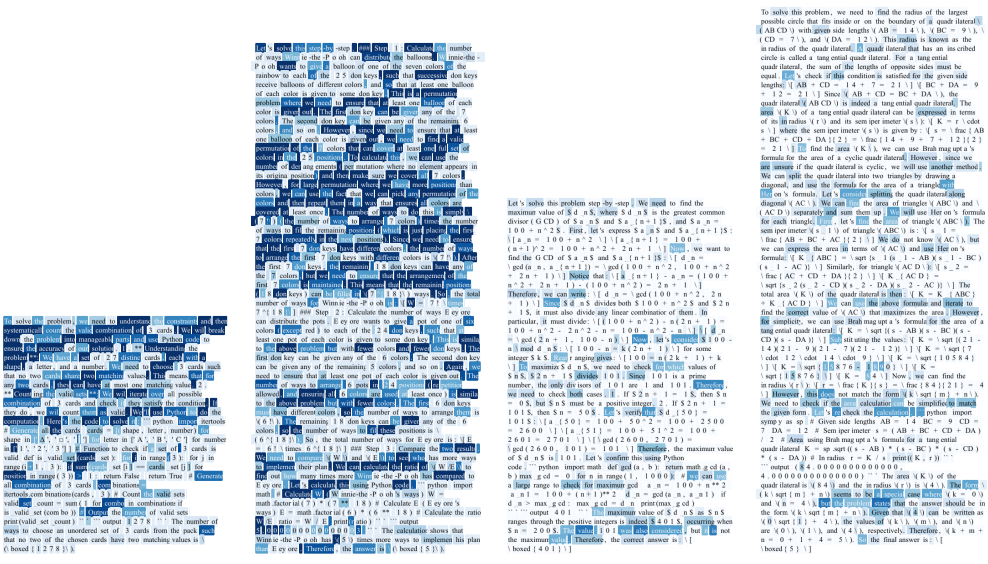
What does relative-entropy bring to the training? Introducing relative entropy into PREPO further enhances training efficiency. As shown in Figure 6, incorporating relative-entropy weighting further reduces the zero-advantage ratio across both Qwen2.5-Math-7B and Qwen2.5-Math-1.5B. This reduction implies improved sample efficiency: a higher fraction of rollouts contributes meaningful learning signals.



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

Figure 6: Comparison of zero advantage ratio between PPL-schedule and PREPO.

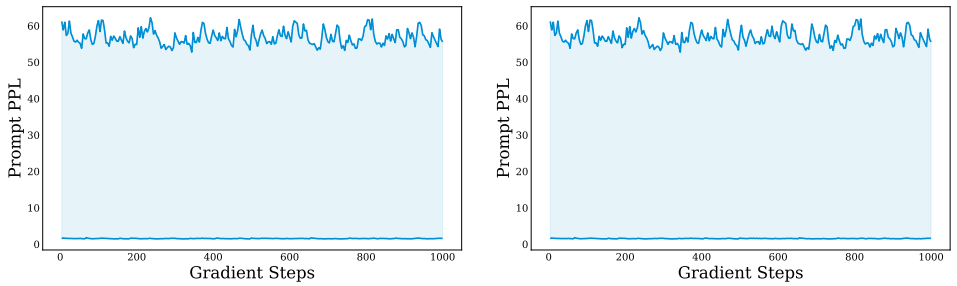
Case Analysis. To visualize the effect of relative entropy weighting, Figure 7 displays examples of token-level entropy from rollouts within the same mini-batch, where darker shades indicate higher token entropy and the red multiplier denotes the relative entropy weight w_i. Rollouts with higher relative entropy (e.g., $\times$ 1.65) correspond to more exploratory reasoning paths, whereas lower-weight rollouts ($\times$ 0.73) reflect more confident, deterministic responses. This analysis highlights how PREPO adaptively balances learning signals that uncertain yet potentially informative rollouts receive stronger gradient influence, while confident but low-diversity outputs are proportionally downweighted.



(a) Correct answer $\times$ 1.65 (b) Incorrect answer $\times$ 1.33 (c) Correct answer $\times$ 0.73 (d) Incorrect answer $\times$ 0.74

Figure 7: Token-level entropy of sequences within a mini-batch (with relative-entropy in red).

Does the prompt PPL vary during training? To examine whether the distribution of prompt difficulty shifts throughout training, we track the range of prompt PPL values at each epoch (Figure 8). The results show that the overall PPL range remains stable, and the mean prompt PPL exhibits minimal drift. This indicates that PREPO's sampling strategy maintains a consistent distribution of task difficulty, preventing a bias toward easier prompts.



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

Figure 8: Range of prompt PPL during training.

Does PREPO affect training time per step? As illustrated in Figure 9, computing prompt PPL introduces only negligible overhead relative to rollout generation time. This confirms that PREPO's filtering and weighting procedures are computationally efficient and thus scalable to larger models and datasets.

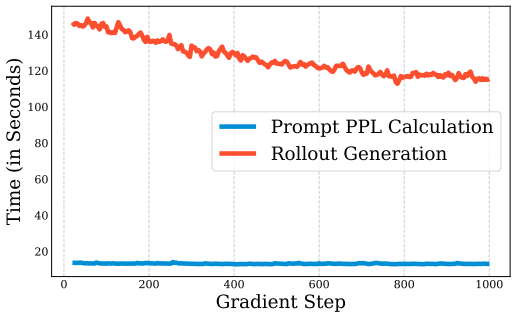


Figure 9: Comparison of calculating prompt PPL and rollout generation

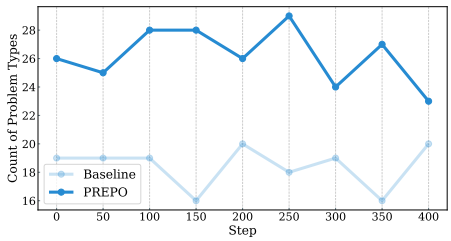
Does PREPO select diverse problems? A further analysis of the training corpus reveals that PREPO samples a more diverse set of mathematical problems than random selection. As shown in Figure 10, PREPO achieves broader coverage across Mathematics Subject Classification (MSC[4]) categories, indicating that its adaptive filtering promotes exposure to a wider range of reasoning types.

Does the model memorize the training data? To evaluate whether PREPO encourages memorization, we follow the methodology of Wu et al. [2025] by truncating 40% of each prompt to create partial problems. Models are then evaluated on these partial inputs, with performance measured by the average pass rate over 16 generations. As shown in Figure 11, most partial problems yield near-zero pass rates, with only a small fraction achieving moderate success. This distribution implies that the models rely on full contextual information to solve tasks rather than recalling memorized solutions. Hence, PREPO's improvements can be attributed to enhanced reasoning and generalization rather than rote memorization of training data.

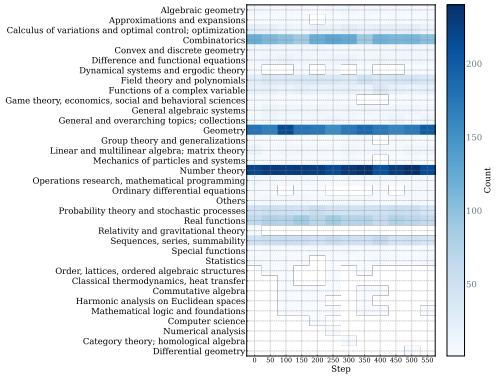
7 Conclusion and Future Work

This study examined how intrinsic data properties can improve the efficiency of RLVR training. On the prompt side, we showed that prompt perplexity provides an effective indicator of model adaptability and naturally supports a data selection scheduling. On the rollout side, sequence-level entropy offered an intrinsic measure of response confidence. Building on these properties to guide

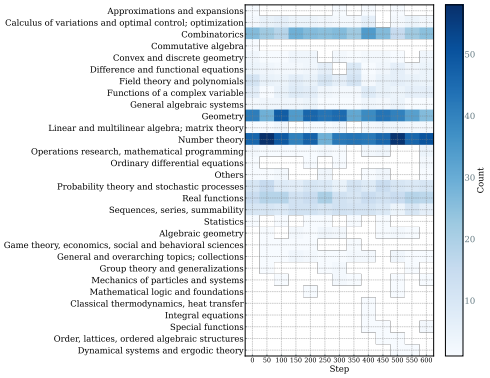
[4]https://zbmath.org/classification/



(a) Unique number of MSC tags

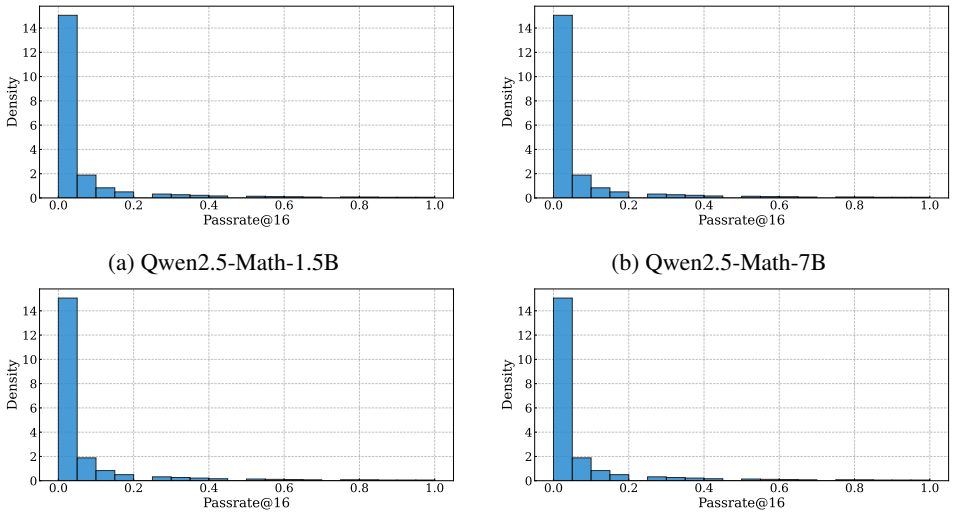


(b) MSC frequency of PREPO



(c) MSC frequency of baseline (random)

Figure 10: Comparison of problem diversity between PREPO and baseline on Qwen2.5-Math-7B.



(a) Qwen2.5-Math-1.5B

(b) Qwen2.5-Math-7B

(c) Qwen3-4B

(d) Llama3.1-8B

Figure 11: Distribution of passrate@16 over all partial prompts

exploration during training, we introduced PREPO, which integrates a perplexity-based schedule with entropy-based weighting. Empirical results from comprehensive experiments shows the effectiveness of PREPO. Future work may extend this direction by exploring additional intrinsic signals and integrating data-driven exploration with system-level optimization methods.

8 Limitations

This study has several limitations that should be acknowledged. (1) The online selection ratio was fixed at 20%, and the impact of alternative ratios has not been systematically examined; (2) response lengths were restricted to 32K tokens, leaving the applicability of PREPO to models generating substantially longer outputs an open question; and (3) the evaluation was limited to mathematical reasoning tasks, while its effectiveness in other domains remains to be explored.

References

Art of Problem Solving. 2024 aime i. https://artofproblemsolving.com/wiki/index.php/2024_AIME_I, 2024. Accessed: 2025-06-09.

Art of Problem Solving. 2025 aime i. https://artofproblemsolving.com/wiki/index.php/2025_AIME_I, 2025. Accessed: 2025-06-09.

Sanghwan Bae, Jiwoo Hong, Min Young Lee, Hanbyul Kim, JeongYeon Nam, and Donghyun Kwak. Online difficulty filtering for reasoning oriented reinforcement learning. *arXiv preprint arXiv:2504.03380*, 2025.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, and Yoshua Bengio. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*, 2025a.

Xinjie Chen, Minpeng Liao, Guoxin Chen, Chengxi Li, and Biao Fu. From data-centric to sample-centric: Enhancing llm reasoning via progressive optimization. *arXiv preprint arXiv:2507.06573*, 2025b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Ehsan Kamalloo, Shipeng Li, Shikun Li, Zhiqin Yang, Xinghua Zhang, Gaode Chen, Xiaobo Xia, Hengyu Liu, and Zhe Peng. Learnalign: Reasoning data selection for reinforcement learning in llms based on improved gradient alignment. *arXiv preprint arXiv:2506.11480*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Haotong Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023. URL https://api.semanticscholar.org/CorpusID:261697361.

Kai Li, Xinggao Fan, and X. Liu. Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning. *arXiv preprint arXiv:2504.13818*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023a.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023b.

X Liu et al. Improving data efficiency for llm reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. *arXiv preprint arXiv:2506.05316*, 2025.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

Yun Qu, Qi Wang, Yixiu Mao, Vincent Tao Hu, Björn Ommer, and Xiangyang Ji. Can prompt difficulty be online predicted for accelerating rl finetuning of reasoning models? *arXiv preprint arXiv:2507.04632*, 2025.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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*, pages 1279–1297, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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

Zhenting Wang and Cui Guofeng. Dump: Automated distribution-level data selection for rl. *arXiv preprint arXiv:2504.09710*, 2025.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Yanwei Fu, Qin Liu, et al. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *arXiv preprint arXiv:2507.10532*, 2025.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Enci Zhang, Xingang Yan, Wei Lin, Tianxiang Zhang, and Qianchun Lu. Learning like humans: Advancing llm reasoning capabilities via adaptive difficulty curriculum learning and expert-guided self-reformulation. *arXiv preprint arXiv:2505.08364*, 2025.

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. *arXiv preprint arXiv:2506.02177*, 2025.

Yinmin Zhong, Zili Zhang, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, et al. Optimizing rlhf training for large language models with stage fusion. *arXiv preprint arXiv:2409.13221*, 2024.

A Entropy as a Confidence Signal in Rollouts

While prompt filtering determines the distribution of prompts x_i presented to the model, standard RLVR methods rely solely on reward feedback and lack an explicit mechanism for assessing the confidence of generated responses. Entropy serves as a common measure of confidence, as it quantifies the sharpness of the predictive distribution along a rollout.

The token-level entropy of a rollout is defined as $H_t = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid o_{<t}, x) \log \pi_\theta(v \mid o_{<t}, x)$, where $\mathcal{V}$ is the vocabulary. The sequence-level entropy is then the average

$$\bar{H}(o \mid x) = \frac{1}{|o|} \sum_{t=1}^{|o|} H_t. \tag{9}$$

Rollouts with low entropy correspond to highly concentrated predictive distributions, reflecting strong model confidence in a single continuation. In contrast, rollouts with high entropy correspond to more diffuse distributions, where multiple continuations remain plausible.

Thus, entropy complements prompt filtering by providing an intrinsic measure of response confidence that can be directly incorporated into training.

B Additional Results on Training Dynamics of Low- and High-PPL Prompts

Across Qwen models (Figure 12 and Figure 13), we observe the following patterns:

- **Entropy.** Prompts with high perplexity (High-PPL) have consistently higher entropy compared to those with low perplexity (Low-PPL).
- **Reward.** Low-PPL prompts receive higher reward values compaired with the High-PPL group.
- **All-Correct Ratio.** Low-PPL prompts reach saturation more rapidly, with a larger proportion of prompts producing fully correct responses.
- **Validation Score (AIME24 [Art of Problem Solving, 2024]).** The Low-PPL group also achieves superior validation performance in the early training stages, suggesting that exposure to more "familiar" data facilitates faster adaptation and knowledge acquisition in large language models.
- **Zero-Advantage Ratio.** In later training phases, the Low-PPL group has a higher zero-variance ratio, resulting in fewer effective rollouts and reduced sample efficiency compared to the High-PPL group.

For Llama3.1-8B (Figure 14), we similarly observe that High-PPL prompts lead to higher entropy and lower reward curves. However, overall performance remains poor because the dataset is too challenging for Llama3.1-8B to be effectively trained with RLVR.

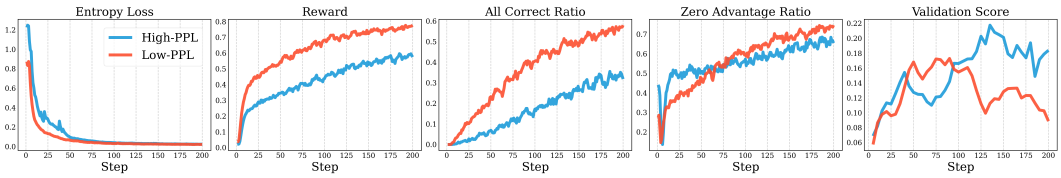


Figure 12: Training dynamics of LOW-PPL vs. HIGH-PPL prompts on Qwen2.5-7B.

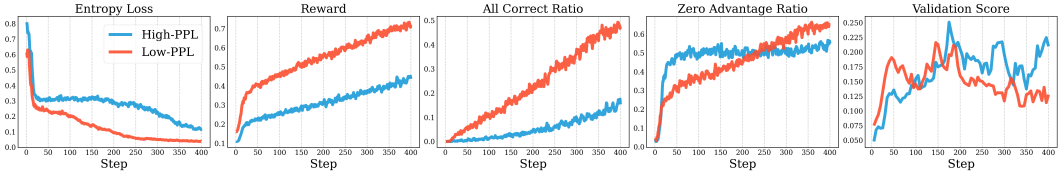


Figure 13: Training dynamics of LOW-PPL vs. HIGH-PPL prompts on Qwen2.5-Math-1.5B.

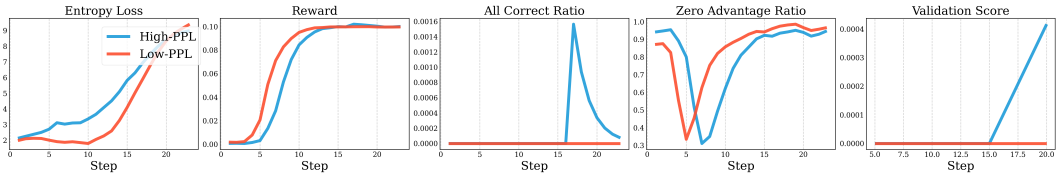
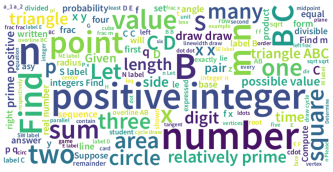


Figure 14: Training dynamics of LOW-PPL vs. HIGH-PPL prompts on Llama3.1-8B.

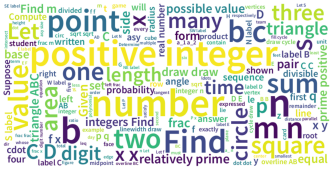
C Visualization of High-PPL and Low-PPL Prompts

As illustrated in Figure 16, High-PPL prompts exhibit a greater prevalence of non-English characters relative to Low-PPL prompts. This pattern is consistently observed across both the Qwen2.5-series and Llama model families.

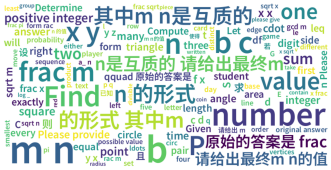


(a) Low-PPL prompts (Qwen2.5-7B)

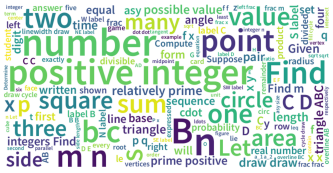
(b) High-PPL prompts (Qwen2.5-7B)



(c) Low-PPL prompts (Qwen2.5-Math-1.5B)



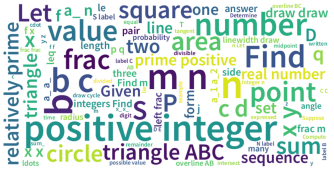
(d) High-PPL prompts (Qwen2.5-Math-1.5B)



(e) Low-PPL prompts (Qwen2.5-Math-7B)

(f) High-PPL prompts (Qwen2.5-Math-7B)

Figure 15: Wordcloud of the most frequent words in Low-/High-PPL prompts



(a) Low-PPL prompts (Llama3.1-8B)

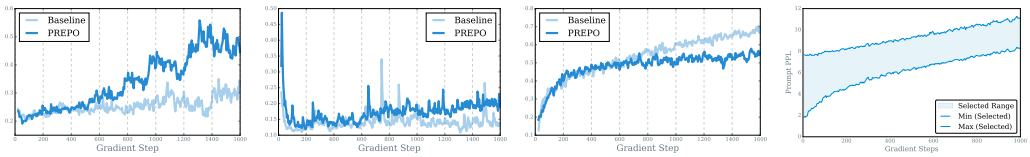
(b) High-PPL prompts (Llama3.1-8B)

Figure 16: Wordcloud of the most frequent words in Low-/High-PPL prompts

D Additional Results

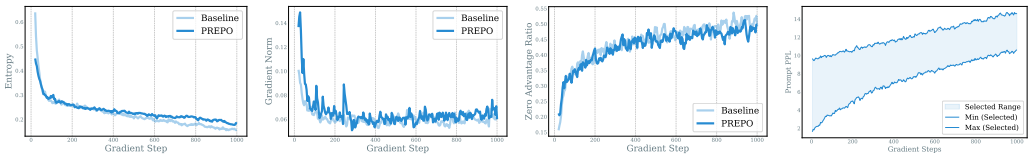
D.1 Training Dynamics of PREPO versus the Baseline

As shown in Figure 17 and 18, both PREPO and the baseline show increasing entropy, though the rise is more pronounced under PREPO, indicating stronger exploration. PREPO also maintains a slightly higher gradient norm without instability. Moreover, it produces a smaller proportion of rollouts with zero advantage, suggesting that its rollouts yield more informative learning signals. Finally, the prompt perplexity selected by PREPO increases gradually throughout training, reflecting a consistent learning progression.



(a) Entropy (b) Gradient norm (c) Zero adv. ratio (d) PREPO PPL range

Figure 17: Full Comparison between PREPO and random selection on Qwen2.5-Math-7B.



(a) Entropy (b) Gradient norm (c) Zero adv. ratio (d) PREPO PPL range

Figure 18: Full Comparison between PREPO and random selection on Qwen2.5-Math-1.5B

D.2 Comparison with No-Filtering

For Qwen2.5-Math-7B, we observe that PREPO attains performance comparable to training on the full dataset without any filtering, i.e., using 5× rollouts per step, as shown in Figure 19

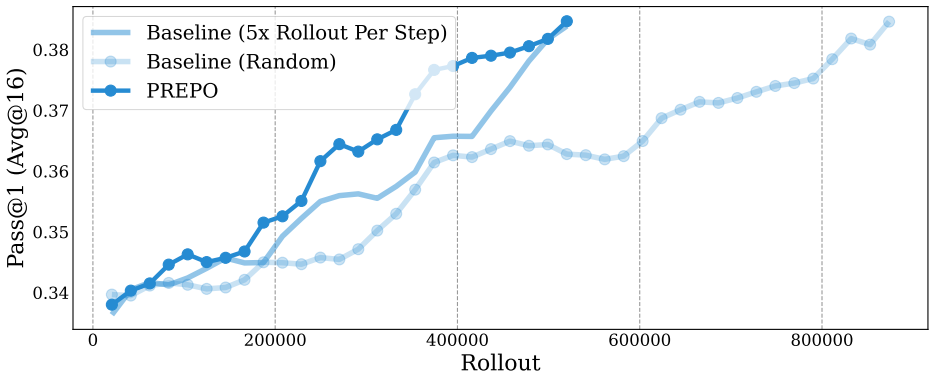


Figure 19: Comparison of PREPO and baselines (a) training w/o filtering (b) random selection.

E Properties of PREPO

Setup. The PREPO objective is defined as

$$\mathcal{J}_{\text{PREPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{I}_\rho,\ \{o_i\}_{i=1}^{G}\sim\pi_{\text{old}}} \left[\frac{1}{G}\sum_{i=1}^{G} w_i \cdot \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} s_{i,t}(\theta)\,\hat{A}_{i,t} \right], \tag{10}$$

where

$$s_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid x, o_{i,<t})}{\pi_{\text{old}}(o_{i,t} \mid x, o_{i,<t})}, \qquad \hat{A}_{i,t} = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^{G})}{\text{std}(\{r_j\}_{j=1}^{G})}.$$

Each rollout i is further scaled by a sequence-level entropy weight

$$\bar{H}_i = \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} H_{i,t}, \qquad \bar{H} = \frac{1}{\sum_{k=1}^{B}|o_k|}\sum_{k=1}^{B}\sum_{t=1}^{|o_k|} H_{k,t}, \qquad w_i = \frac{\bar{H}_i}{\bar{H}},$$

where B denotes the micro-batch size used for a single gradient update.

E.1 Sum of weights vs. batch size

The weights are normalized relative to $\bar{H}$, so that

$$\frac{1}{B}\sum_{i=1}^{B} w_i \cdot |o_i| = \frac{1}{B\bar{H}}\sum_{i=1}^{B}|o_i|\bar{H}_i = \frac{1}{B\bar{H}}\sum_{i=1}^{B}\sum_{t=1}^{|o_i|} H_{i,t} = \frac{\sum_{i=1}^{B}|o_i|}{B}. \tag{11}$$

Thus the *token-weighted average weight equals the average sequence length*. If all sequences have equal length, then $\frac{1}{B}\sum_i w_i = 1$.

As shown in Figure 20, the effective batch size remains close to the nominal batch size throughout training. In our experiments, it starts at 1.04 and gradually declines to 0.98 during RLVR training. Early in training, when low-PPL prompts dominate and produce confident (low-entropy) responses, the average effective weight exceeds 1, reflecting a relative emphasis on the few higher-entropy rollouts. As training progresses and higher-PPL prompts enter, the overall entropy distribution shifts upward. After normalization, this causes the average effective weight to fall slightly below 1.

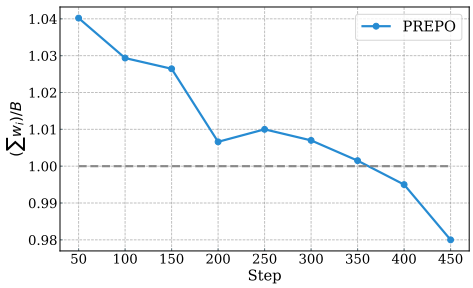


Figure 20: Trend of the effective batch size (Qwen2.5-Math-1.5B).

E.2 Sensitivity to extreme entropies

Because $w_i = \bar{H}_i/\bar{H}$ is normalized by the batch mean, a rollout with extremely large entropy $\bar{H}_j \gg \bar{H}$ does not inflate its own weight ($w_j \approx 1$), but instead suppresses the weights of other rollouts ($w_i \ll 1$ for $i \neq j$). The partial derivative of the relative-entropy weight with respect to a sequence-level entropy in the batch is

$$\frac{\partial w_i}{\partial \bar{H}_j} = \begin{cases} \dfrac{1}{\bar{H}} - \dfrac{\bar{H}_j}{\bar{H}^2}\dfrac{|o_j|}{\sum_k |o_k|}, & i = j, \\ -\dfrac{\bar{H}_i}{\bar{H}^2}\dfrac{|o_j|}{\sum_k |o_k|}, & i \neq j, \end{cases}$$

which shows that (1) for the self-sensitivity ($i = j$), the two terms nearly cancel, so increasing $\bar{H}_j$ has only a small net effect on w_j itself; (2) for cross-sensitivity ($i \neq j$), the derivative is strictly negative, meaning that increasing $\bar{H}_j$ decreases the weights of all other rollouts.

Thus, extreme entropies influence the distribution of weights not by amplifying the outlier's own contribution, but by enlarging the batch mean $\bar{H}$, which in turn uniformly shrinks the normalized weights of the remaining rollouts in proportion to their entropies.

As shown in Figure 21, the distribution of relative-entropy weights is sharply concentrated around 1, with most values between 0.7 and 1.5. A small fraction of rollouts appear in the long tail beyond 2, with rare outliers above 4. This pattern is consistent with the analysis above, that is, entropy outliers remain bounded in their own weights but shift the normalization, thereby downscaling the majority of rollouts.

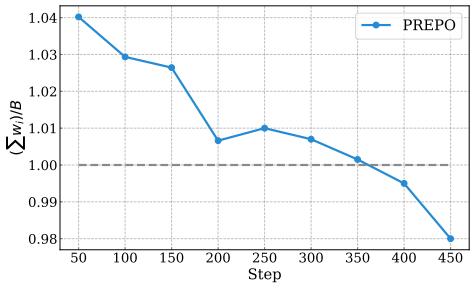


Figure 21: Frequency of relative-entropy weights (Qwen2.5-Math-1.5B; 50 training step)

Remark on clipping. For clarity, clipping was omitted from the above derivation. In practice, PPO-style clipping only truncates extreme importance ratios $s_{i,t}(\theta)$, effectively setting the corresponding gradient contributions to zero.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims are summarized in Figure 1 and Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include the limitations of our work in Section 8

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is not a theoretical paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explained our settings in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Github link provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are summarized in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We show the average performance of over multiple runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Details are summarized in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the understood the code of ethics; and have done our best to conform.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work does not impact the society at large.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Since we use open-source models and datasets, our work poses no risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited open-sourced libraries in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release our code base with included readme files.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines: This work does not involve crowdsourcing nor research within human subjects.

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research within human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: In this work, LLMs were employed solely for grammar correction and sentence rephrasing. They had no involvement in research ideation, experimental design, data analysis, or substantive writing. Their role was restricted to improving clarity and style; therefore, they are not considered contributors to the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.