
Supplementary Material for Paper #13557

"Act Only When It Pays: Efficient Reinforcement Learning for LLM Reasoning via Selective Rollouts"

1 **A Overview**

2 We begin in Section B by discussing the limitations of our method. Section C highlights the broader
3 societal and practical impact of improving rollout efficiency for LLM training. Section E details our
4 experimental setup, and Section F presents additional empirical experiments and analysis.

5 **B Limitations**

6 While GRESO effectively filters out the most obvious zero-variance training prompts—those that
7 contribute no learning signal to the model, it does not estimate or rank the value of the remaining
8 prompts, which can also contain uninformative prompts that provide limited contribution to training.
9 A potential future work for GRESO is to extend its filtering mechanism beyond binary decisions by
10 incorporating a finer-grained scoring or ranking system to prioritize prompts based on their estimated
11 training utility. Despite that, we view GRESO as an important first step toward such an advanced data
12 selection algorithm for efficient rollout and believe it provides a solid foundation for more adaptive
13 and efficient reinforcement learning in LLM training.

14 **C Broader Impact**

15 This work enhances the efficiency and scalability of RL-based fine-tuning for language models by
16 introducing a lightweight, selective rollout mechanism that filters out uninformative prompts. By
17 significantly reducing redundant computation, our method lowers overall training costs. This makes it
18 easier for institutions with limited computational budgets to train strong models, helping democratize
19 access to advanced AI. Furthermore, our approach promotes more sustainable and resource-efficient
20 practices, encouraging future research toward greener and more inclusive large-scale training.

21 **D Reproductivity**

22 We introduced our detailed experimental setting in Section E, and we also include our code in the
23 supplementary material.

24 **E Detailed Experimental Setting**

25 **Models & Datasets.** We run our experiments on Qwen2.5-Math-1.5B [14], DeepSeek-R1-Distill-
26 Qwen-1.5B [3], and Qwen2.5-Math-7B [14]. For Qwen2.5-Math-1.5B/7B models, we use 4096 as the
27 context length, as it is the maximum context length for those two models. For DeepSeek-R1-Distill-
28 Qwen-1.5B, we set the context length to 8196. For training datasets, we train our methods on two
29 datasets in two settings: 1) DAPO+MATH (DM): We combine the DAPO dataset [15], which contains
30 only integer solutions, with the MATH dataset [6], which also contains LaTeX-formatted solutions.
31 We find that training on DAPO alone can degrade performance on LaTeX-based benchmarks, so we
32 augment it with MATH to preserve formatting diversity and improve generalization. 2) OPEN-R1
33 30k subset (R1): A 30,000-example subset of the OPEN-R1 math dataset [4].

Training. Our method is implemented based on verl [12] pipeline and uses vLLM [8] for rollout. We use 4xH100 for Qwen2.5-Math-1.5B training and 8xH100 for Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-1.5B. We set the rollout temperature to 1 for vLLM [8]. The training batch size is set to 256, and the mini-batch size to 512. We sample 8 responses per prompt. We set the default rollout sampling batch size as 384. For DeepSeek-R1-Distill-Qwen-1.5B, we set the context length to 8196. The training batch size is set to 128, and the mini-batch size to 512. We also sample 8 responses per prompt. We set the default rollout sampling batch size as 192. We train all models for 1000 steps, and we optimize the actor model using the AdamW [11] optimizer with a constant learning rate of 1e-6. We use $\beta_1 = 0.9$, $\beta_2 = 0.999$, and apply a weight decay of 0.01. We use the following question template to prompt the LLM. For reward assignment, we give a score of 0.1 for successfully extracting an answer and a score of 1.0 if the extracted answer is correct. Similar to [15], we remove the KL-divergence term. The optimization is performed on the parameters of the actor module wrapped with Fully Sharded Data Parallel (FSDP) [17] for efficient distributed training. We use 4 H100 for Qwen2.5-Math-1.5B training and 8 H100 for Qwen2.5-Math-7B and DeepSeek-R1-Distill-Qwen-1.5B (as it has a longer context length.) We set the targeted zero-variance percentage to 25% for all experiments and allocate it between easy and hard prompts in an 1 : 2 ratio (i.e., 8.3% for easy and 16.7% for hard zero-variance prompts), based on the intuition that, as models become more capable during training, more exploration on hard examples can be more beneficial. However, a more optimal allocation scheme may exist, which we leave for future study. We set the initial exploration probability to 50% and base exploration probability adjustment step size Δp for base exploration probability to 1%. We also set a minimal base exploration probability to 5% to ensure a minimal level of exploration on zero-variance prompts throughout training.

GRESO with Fixed Parameters Across All Experiments. Although GRESO introduces a few hyperparameters, we argue that hyperparameter tuning is not a major concern in practice. We designed GRESO (e.g., self-adjustable base exploration probability) to be robust under default settings and conducted all experiments using a single fixed set of hyperparameters across models and tasks. The consistent performance observed across different models and tasks demonstrates that GRESO does not rely on extensive hyperparameter tuning, making it both practical and easy to integrate into existing RL fine-tuning pipelines.

Evaluation. For benchmark datasets, we use six widely used complex mathematical reasoning benchmarks to evaluate the performance of trained models: Math500 [6, 10], AIME24 [1], AMC [2], Minerva Math [9], Gaokao [16], Olympiad Bench [5]. Same as the training setting, For Qwen2.5-Math-1.5B/7B models, we use 4096 as the context length. For DeepSeek-R1-Distill-Qwen-1.5B, we set the context length to 8196. Similar to [13], we evaluate models on those benchmarks every 50 steps and report the performance of the checkpoint that obtains the best average performance on six benchmarks. We evaluate all models with temperature = 1 and repeat the test set 4 times for evaluation stability, i.e., $pass@1(avg@4)$, for all benchmarks.

Question Template

Please solve the following math problem: {{Question Description}}. The assistant first thinks about the reasoning process step by step and then provides the user with the answer. Return the final answer in \boxed{} tags, for example \boxed{1}. Let's solve this step by step.

71

72 F Additional Experiments

73 F.1 Impact of Targeted Zero-variance Percentage

74 In this section, we study how varying the targeted zero-variance
 75 percentage impacts training and rollout efficiency. In addition
 76 to the default setting of 25% used throughout our experiments,
 77 we also evaluate alternative values of 0, 50%, 100% (i.e., always
 78 allow exploration). As shown in Table 1, different zero-variance
 79 targets give us nearly identical accuracy. We also present the
 80 number of rollouts per step in Figure 1. When we reduce the
 81 targeted zero-variance ratio to 0, we observe that the number
 82 of rollouts per step remains similar to that of the 25% setting.
 83 This lack of difference can be attributed to two factors. First, we
 84 enforce a minimum exploration rate of 5%, which ensures that
 85 some exploration still occurs. As a result, the actual zero-variance
 86 percentage never truly reaches 0. Second, we always oversample
 87 some data in the first batch of rollouts in each iteration to provide
 88 some redundancy to avoid the second batch of rollouts. With this setting, as long as the first batch
 89 generates enough effective training data to fill the training batch, regardless of whether the target is 0
 90 or 25%, the total number of rollouts remains approximately the same. In addition, as the targeted
 91 zero-variance percentage increases, more zero-variance prompts are allowed during rollout, leading
 92 to a higher number of rollouts per step. When the targeted percentage becomes sufficiently large,
 93 GRESO gradually approaches the behavior of dynamic sampling with adaptive rollout batch size.

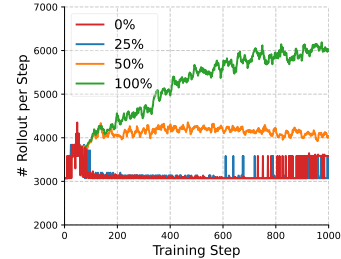


Figure 1: Comparison of the number of rollouts across different target zero-variance ratios.

Table 1: Average accuracy across six math reasoning benchmarks under different targeted zero-variance percentages.

Target (%)	0	25	50	100
Acc. (%)	48.1	48.5	48.5	48.4

94 F.2 Alternative Design: Linear Backoff

95 In addition to the probabilistic filtering approach introduced in
 96 Section 4.2 of the main paper, we also explored an alternative
 97 solution for filtering zero-variance prompts during the early stages
 98 of this project. One such method is the *backoff algorithm* [7]
 99 (e.g., linear backoff). Specifically, if a prompt is identified as
 100 zero-variance in the most recent k rollouts, it is skipped for the
 101 next k training epochs. However, there are several limitations to
 102 this approach. As discussed in Section 4 of the paper, the degree
 103 of exploration should adapt to the model, dataset, and training
 104 stage. The linear backoff algorithm schedules the next rollout
 105 for a zero-variance prompt k epochs into the future. As a result,
 106 if we wish to adjust the exploration intensity dynamically based
 107 on new observations or evolving training dynamics, the backoff
 108 algorithm cannot directly affect prompts that have already been
 109 deferred to future epochs. For instance, as shown in Figure 2, unlike probabilistic filtering, filtering
 110 based on linear backoff can cause periodic fluctuations in zero-variance prompt ratio, which differs
 111 from the smoother dynamics enabled by probabilistic filtering. This lack of flexibility limits its ability
 112 to adapt exploration strategies in a fine-grained or responsive manner, which motivated the design of
 113 our current GRESO algorithm based on probabilistic filtering.

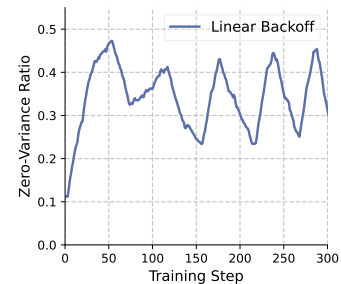


Figure 2: Zero-variance prompt ratio dynamic for linear backoff.

114 F.3 Case study of Filtered Examples

115 To better understand the behavioral patterns of our selective filtering algorithm, we present a case
 116 study of prompts that were frequently skipped or selected during training from the MATH [6] dataset.
 117 We categorize the examples into three groups: Frequently Skipped Prompts (Easy), Frequently
 118 Skipped Prompts (Hard), Frequently Selected Prompts. We observe that frequently skipped easy

119 prompts often involve straightforward calculations or routine applications of formulas, making them
 120 more likely to be solved across all sampled responses. Frequently selected prompts tend to exhibit
 121 moderate difficulty, contributing more consistently to model improvement. As for frequently skipped
 122 hard prompts, these problems are too challenging for the model to solve, even across multiple rollouts,
 123 resulting in zero variance among the rewards and ultimately failing to contribute to training.

Frequently Skipped Prompts (Easy)

1. **Question:** Johnny has 7 different colored marbles in his bag. In how many ways can he choose three different marbles from his bag to play a game? **Solution:** 35.
2. **Question:** The number n is a prime number between 20 and 30. If you divide n by 8, the remainder is 5. What is the value of n ? **Solution:** 29.
3. **Question:** Evaluate: $\frac{10^{-2} \cdot 5^0}{10^{-3}}$ **Solution:** 10.
4. **Question:** The Ponde family's Powerjet pumps 420 gallons of water per hour. At this rate, how many gallons of water will it pump in 45 minutes? **Solution:** 315.
5. **Question:** Suppose that $n, n+1, n+2, n+3, n+4$ are five consecutive integers. Determine a simplified expression for the sum of these five consecutive integers. **Solution:** $5n + 10$.

124

Frequently Skipped Prompts (Hard)

1. **Question:** A parabola and an ellipse share a focus, and the directrix of the parabola is the line containing the minor axis of the ellipse. The parabola and ellipse intersect at two points. Given that the equation of the ellipse is $\frac{x^2}{25} + \frac{y^2}{9} = 1$, find the distance between those two points. **Solution:** $\frac{4\sqrt{14}}{3}$.
2. **Question:** In triangle ABC , $AB = AC = 100$, and $BC = 56$. Circle P has radius 16 and is tangent to \overline{AC} and \overline{BC} . Circle Q is externally tangent to P and is tangent to \overline{AB} and \overline{BC} . No point of circle Q lies outside of $\triangle ABC$. The radius of circle Q can be expressed in the form $m - n\sqrt{k}$, where m, n , and k are positive integers and k is the product of distinct primes. Find $m + nk$. **Solution:** 254.
3. **Question:** Let $EFGH$, $EFDC$, and $EHBC$ be three adjacent square faces of a cube, for which $EC = 8$, and let A be the eighth vertex of the cube. Let I, J , and K , be the points on \overline{EF} , \overline{EH} , and \overline{EC} , respectively, so that $EI = EJ = EK = 2$. A solid S is obtained by drilling a tunnel through the cube. The sides of the tunnel are planes parallel to \overline{AE} , and containing the edges \overline{IJ} , \overline{JK} , and \overline{KI} . The surface area of S , including the walls of the tunnel, is $m + n\sqrt{p}$, where m, n , and p are positive integers and p is not divisible by the square of any prime. Find $m + n + p$. **Solution:** 417.
4. **Question:** Let a and b be nonnegative real numbers such that

$$\sin(ax + b) = \sin 29x$$
 for all integers x . Find the smallest possible value of a . **Solution:** $10\pi - 29$.
5. **Question:** Four people sit around a circular table, and each person will roll a standard six-sided die. What is the probability that no two people sitting next to each other will roll the same number after they each roll the die once? Express your answer as a common fraction. **Solution:** $\frac{35}{72}$.

125

Frequently Selected Prompts

1. **Question:** Let x, y , and z be three positive real numbers whose sum is 1. If no one of these numbers is more than twice any other, then find the minimum value of the product xyz .

Solution: $\frac{1}{32}$.

2. **Question:** The number

$$e^{7\pi i/60} + e^{17\pi i/60} + e^{27\pi i/60} + e^{37\pi i/60} + e^{47\pi i/60}$$

is expressed in the form $re^{i\theta}$, where $0 \leq \theta < 2\pi$. Find θ . **Solution:** $\frac{9\pi}{20}$.

3. **Question:** For what values of x is

$$\frac{x - 10x^2 + 25x^3}{8 - x^3}$$

nonnegative? Answer as an interval. **Solution:** $[0, 2)$.

4. **Question:** Determine all real numbers a such that the inequality $|x^2 + 2ax + 3a| \leq 2$ has exactly one solution in x . **Solution:** 1, 2.

5. **Question:** By starting with a million and alternatively dividing by 2 and multiplying by 5, Anisha created a sequence of integers that starts 1000000, 500000, 2500000, 1250000, and so on. What is the last integer in her sequence? Express your answer in the form a^b , where a and b are positive integers and a is as small as possible. **Solution:** 5^{12} .

References

- [1] Art of Problem Solving. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2025-04-20.
- [2] Art of Problem Solving. Amc problems and solutions. https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions. Accessed: 2025-04-20.
- [3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qishui Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [4] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025.
- [5] 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.
- [6] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- [7] Byung-Jae Kwak, Nah-Oak Song, and Leonard E Miller. Performance analysis of exponential backoff. *IEEE/ACM transactions on networking*, 13(2):343–355, 2005.
- [8] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.
- [9] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.
- [10] 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*, 2023.
- [11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.
- [12] 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. *arXiv preprint arXiv: 2409.19256*, 2024.

- 183 [13] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He,
184 Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one
185 training example. *arXiv preprint arXiv:2504.20571*, 2025.
- 186 [14] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu,
187 Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via
188 self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.
- 189 [15] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu,
190 Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv*
191 *preprint arXiv:2503.14476*, 2025.
- 192 [16] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the
193 performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.
- 194 [17] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid
195 Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel.
196 *arXiv preprint arXiv:2304.11277*, 2023.