
Single-stream Policy Optimization

Zhongwen Xu* and Zihan Ding*

Tencent

Abstract

We revisit policy-gradient optimization for Large Language Models (LLMs) from a single-stream perspective. Prevailing group-based methods like GRPO reduce variance with on-the-fly baselines but suffer from critical flaws: frequent degenerate groups erase learning signals, and synchronization barriers hinder scalability. We introduce Single-stream Policy Optimization (SPO), which eliminates these issues by design. SPO replaces per-group baselines with a persistent, KL-adaptive value tracker and normalizes advantages globally across the batch, providing a stable, low-variance learning signal for every sample. Being group-free, SPO enables higher throughput and scales effectively in long-horizon or tool-integrated settings where generation times vary. Furthermore, the persistent value tracker naturally enables an adaptive curriculum via prioritized sampling. Experiments using Qwen3-8B show that SPO converges more smoothly and attains higher accuracy than GRPO, while eliminating computation wasted on degenerate groups. Ablation studies confirm that SPO’s gains stem from its principled approach to baseline estimation and advantage normalization, offering a more robust and efficient path for LLM reasoning. Across five hard math benchmarks with Qwen3-8B, SPO improves the average maj@32 by +3.4 percentage points (pp) over GRPO, driven by substantial absolute point gains on challenging datasets, including +7.3 pp on BRUMO 25, +4.4 pp on AIME 25, +3.3 pp on HMMT 25, and achieves consistent relative gain in pass@ k across the evaluated k values. SPO’s success challenges the prevailing trend of adding incidental complexity to RL algorithms, highlighting a path where fundamental principles, not architectural workarounds, drive the next wave of progress in LLM reasoning.

1 Introduction

Reinforcement learning (RL) [36] has become a cornerstone for advancing the reasoning capabilities of Large Language Models (LLMs), notably the Reinforcement Learning with Verifiable Reward (RLVR) paradigm [21, 11]. Methods like Group Relative Policy Optimization (GRPO) [34, 11] have achieved remarkable success by adopting a *multi-outcome* approach, generating a group of responses for each prompt to construct an on-the-fly baseline for variance reduction. While this “group-based” paradigm has pushed the state of the art, it suffers from fundamental inefficiencies. When all responses in a group share the same outcome (e.g., all correct or all incorrect), the relative advantage collapses to zero, yielding no learning signal. This degeneracy represents a fundamental waste of computation and data. To counteract this, a series of engineering heuristics like dynamic sampling [42] have been developed. These workarounds, while functional, add significant complexity and create a less principled, more convoluted optimization process.

Group-based architectural choice also imposes a critical synchronization barrier. In distributed training, the entire group must wait for its slowest member, a bottleneck that becomes particularly acute in complex agentic tasks requiring multi-turn tool use or long-horizon reasoning [15, 41, 45].

*Equal contributions, corresponding to: zhongwenxu@tencent.com, dingzihan737@gmail.com

In these settings, interaction times are highly variable (e.g., number of interaction turns, time per interaction, etc), and a single slow-running agentic trajectory can stall its entire group, severely hindering training throughput and scalability.

We advocate for returning to the classic single-stream paradigm for policy gradient optimization [36], where each training sample is a single stream of prompt-response pair. This is not a mere simplification, but a deliberate re-alignment with foundational RL principles to address the aforementioned architectural flaws. To overcome the critical challenge of high gradient variance in this setting, we introduce Single-stream Policy Optimization (SPO). SPO replaces the noisy, on-the-fly group baseline with three synergistic components for stable and efficient learning. First, it employs a lightweight Bayesian value tracker to maintain a persistent, temporally-informed estimate of the success probability for each prompt, serving as a low-variance baseline. Second, it normalizes advantages globally across the entire batch, avoiding the instability of per-group statistics. Finally, this architecture naturally enables an adaptive curriculum via prioritized sampling, focusing computational resources on the most informative prompts.

The benefits of this principled approach are clear: SPO is inherently more scalable and eliminates the computational waste of degenerate groups. Our experiments confirm these advantages, demonstrating that SPO consistently outperforms GRPO on challenging reasoning benchmarks, improving the absolute point gains on challenging datasets, including 7.3 percentage points (pp) on BRUMO 25, 4.4 pp on AIME 25, 3.3 pp on HMMT 25, and the pass@ k curves of SPO are above GRPO for all k s. The scalability benefit is particularly pronounced in agentic settings. Our simulations, designed to model these variable-time scenarios, show that SPO’s group-free design can achieve a $4.35\times$ training throughput speedup by eliminating group synchronization bottlenecks. SPO thus provides a more robust foundation for modern LLM optimization, prompting a re-evaluation of essential versus incidental complexity in the field.

2 Method

We introduce Single-stream Policy Optimization (SPO), a method designed for policy optimization in settings with verifiable feedback (RLVR) [21]. We assume the feedback is binary², i.e., +1 for success and 0 for failure. SPO addresses the challenge of estimating a non-stationary success probability for a policy that evolves over training iterations. It integrates a Bayesian value tracker with an adaptive memory mechanism into a policy gradient framework. The core components are: (1) a KL-adaptive tracker that provides a low-variance, single-sample estimate of the success probability; (2) a global advantage normalization scheme that ensures high sample efficiency and stable learning dynamics; and (3) prioritized sampling across training prompts to focus on prompts with high learning potential. The following subsections detail each component.

2.1 A KL-Adaptive Value Tracker

The definition of a value function is the *expected* reward of the prompt x under policy π , i.e., $V_\pi(x) = \mathbb{E}_{y \sim \pi(\cdot|x)}[R(x, y)]$. We use $\hat{v}(x)$ to denote the tracker’s running estimate of $V_\pi(x)$; that is, $\hat{v}(x) \approx V_\pi(x)$. To estimate the non-stationary success probability of a prompt x , we use a Bayesian *tabular* tracker instead of a separate value network³. For the binary success/failure rewards common in RLVR, this is elegantly modeled using a Beta distribution, which is the conjugate prior for the Bernoulli process governing the outcomes. We therefore model the success probability $\hat{v}(x)$ using a Beta distribution: $\hat{v}(x) \sim \text{Beta}(\alpha(x), \beta(x))$, where the value estimate is the posterior mean $\hat{v}(x) = \alpha(x)/(\alpha(x) + \beta(x))$.

The tracker adapts to policy changes by dynamically adjusting its memory of past rewards. When the policy changes significantly, older observations become less relevant and should be downweighted. After each new observation $r(x, y) \in \{0, 1\}$, we discount the prior Beta parameters $(\alpha_{-1}, \beta_{-1})$ by a factor $\rho(x)$ before incorporating the new evidence $r(x, y)$:

$$\alpha(x) = \rho(x)\alpha_{-1}(x) + r(x, y), \quad \beta(x) = \rho(x)\beta_{-1}(x) + (1 - r(x, y)), \quad \hat{v}(x) = \frac{\alpha(x)}{\alpha(x) + \beta(x)}. \quad (1)$$

²Generalizing to non-binary rewards is straightforward, as discussed at the end of Section 2.1.

³The development of core RL algorithms was on tabular representation [36].

The discount factor $\rho(x) = 2^{-D(x)/D_{\text{half}}}$ is determined by the KL divergence $D(x)$ between the current policy and the last policy that *acted* on prompt x , causing the tracker to forget faster as the policy changes more significantly. The hyperparameter D_{half} controls this forgetting rate $\rho \in [\rho_{\min}, \rho_{\max}]$.

Initialization. To initialize, we collect n_0 samples to compute an initial value estimate $\hat{v}_0(x)$. To avoid transient instability, we set the initial effective sample size to its expected equilibrium, $N_0 = 1/(1 - \rho_{\min})$, where ρ_{\min} is the minimum allowed forgetting factor. The initial parameters are then:

$$\alpha_0(x) = N_0 \cdot \hat{v}_0(x), \quad \beta_0(x) = N_0 \cdot (1 - \hat{v}_0(x)). \quad (2)$$

This Bayesian update is equivalent to an adaptive Exponential Moving Average (EMA) on the value estimate:

$$\hat{v}(x) = \hat{v}_{-1}(x) + \eta(x)(r(x, y) - \hat{v}_{-1}(x)), \quad (3)$$

where the learning rate $\eta(x) = (\rho(x)N_{\text{eff},-1}(x) + 1)^{-1}$ naturally adapts to both policy shifts (via $\rho(x)$) and statistical confidence (via $N_{\text{eff}} = \alpha(x) + \beta(x) + 1$). This formulation highlights how our tracker balances new evidence against accumulated knowledge. For *general rewards* beyond binary ones, we can just use the same EMA formulation to directly track \hat{v} , rather than relying on α and β in the binary cases.

2.2 Advantage Estimation and Policy Optimization

SPO uses the tracker’s estimate \hat{v} as a baseline for advantage calculation in a policy gradient algorithm. At iteration i , for a single reward $r(x, y)$ obtained with policy π_{θ_i} , the advantage is computed using the *pre-update* baseline (denoted with subscript $_{-1}$):

$$A(x, y) = r(x, y) - \hat{v}_{-1}(x). \quad (4)$$

Using the baseline from the previous step ensures that it is independent of the action taken at step i , preserving the unbiasedness of the policy gradient estimate. While the reward $r(x, y)$ is typically a direct outcome signal, SPO’s framework is also compatible with more sophisticated reward functions. For instance, recent work like InfAlign [4] demonstrates how to calibrate and transform the reward signal to be “inference-aware,” directly optimizing for procedures like Best-of- N sampling. Such transformed rewards can be seamlessly integrated into SPO by replacing the standard $r(x, y)$ in the advantage calculation. Since $v_{-1}(x)$ is independent of $y \sim \pi_{\theta_i}(\cdot|x)$, $\mathbb{E}[(r - v_{i-1}(x)) \nabla_{\theta} \log \pi] = \nabla J(\theta)$ [39]. Instead of normalizing advantages on a per-prompt basis in a group [34, 11], SPO normalizes them across an entire batch of prompts \mathcal{B} [19, 31, 3, 23]. The normalized advantage $\tilde{A}(x, y)$ is computed as:

$$\tilde{A}(x, y) = \frac{A(x, y) - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}}, \quad (5)$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ are the mean and standard deviation of advantages in the batch $\{A(x, y)\}_{x \in \mathcal{B}}$. We then apply the advantage $\tilde{A}(x, y)$ to each *token* in the response sequence y and update the policy parameters using a standard PPO-Clip policy loss [31]⁴:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_{s,t} \left[\min \left(\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)} \tilde{A}_t, \text{clip} \left(\frac{\pi_{\theta}(a_t | s_t)}{\pi_{\theta_{\text{old}}}(a_t | s_t)}, 1 - \varepsilon, 1 + \varepsilon \right) \tilde{A}_t \right) \right]. \quad (6)$$

Methods like Clip-Higher [42], Clip-Cov [10] and KL-Cov [10] to retain policy entropy are applicable here. Other policy optimization algorithms like CISPO [27] (similar to *vtrace* [12, 40]) and GSPO [43] (use sequence-level likelihood instead of token-level) are compatible with our advantage estimator. Advanced methods to control policy behaviors like ASPO [22] can be utilized to modulate the advantage values. We note that if we use “no baseline” (i.e., $\hat{v} = 0$), it is an extremely simple and valid algorithm but may suffer from high policy gradient variance.

2.3 Prioritized Prompt Sampling

⁴The term “PPO” is frequently used with ambiguity. It may denote the entire algorithm suite (e.g., clipped policy and value losses), refer narrowly to just the clipped policy objective, or describe the broader training framework, including mechanisms like mini-batch updates.

Algorithm 1 Single-stream Policy Optimization

```
1: for iteration  $i = 1, 2, \dots, T$  do
2:   For each  $x \in \mathcal{X}$ , compute sampling weight  $w_i(x)$  according to Eqn. (7).
3:   Sample a batch of  $B$  prompts  $\mathcal{B}_i \subset \mathcal{X}$  according to weights  $\{w_i(x)\}$ .
4:    $\mathcal{D} \leftarrow \emptyset$ 
5:   for each prompt  $x \in \mathcal{B}_i$  do
6:     Sample action  $y \sim \pi_{\theta_{i-1}}(\cdot | x)$  and observe reward  $r(x, y) \in \{0, 1\}$ .
7:     Compute raw advantage  $A(x, y) \leftarrow r(x, y) - \hat{v}_{i-1}(x)$ .
8:     Store  $(x, y, A(x, y))$  in  $\mathcal{D}$ .
9:     Update tracker  $\hat{v}(x)$ .
10:  Normalize advantages:  $\tilde{A}(x, y) \leftarrow (A(x, y) - \mu_{\mathcal{B}_i}) / \sigma_{\mathcal{B}_i}$ .
11:  Update  $\theta_{i-1}$  to  $\theta_i$  using mini-batches with a policy gradient algorithm (e.g., PPO-Clip).
```

To further enhance data efficiency, SPO employs a curriculum learning strategy by prioritizing prompts with the highest learning potential [30, 36]. At each iteration, we sample a batch of prompts based on a score that emphasizes prompts with high uncertainty, while ensuring a minimum level of exploration. The sampling weight $w_i(x)$ for prompt x is defined as:

$$w_i(x) \propto \sqrt{\hat{v}_{i-1}(x)(1 - \hat{v}_{i-1}(x))} + \epsilon. \quad (7)$$

The first term corresponds to the estimated standard deviation of a Bernoulli outcome, which naturally allocates more weight to prompts that are neither almost always solved ($\hat{v} \approx 1$) nor almost always failed ($\hat{v} \approx 0$). The exploration bonus ϵ , set to 0.05 by default, prevents curriculum collapse by ensuring that every prompt retains a non-zero probability of being sampled, thereby maintaining broad coverage of the data distribution. The complete SPO training procedure is outlined in Algorithm 1.

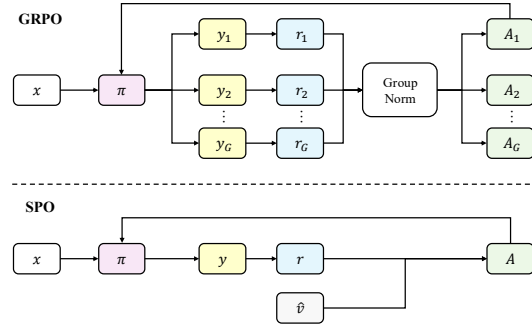


Figure 1: Illustrations of GRPO and SPO.

2.4 Advantages over GRPO

Group-Free for Scalable Infrastructure. SPO’s design is inherently “group-free”, a significant advantage in distributed training frameworks for LLMs. Each sample, consisting of a single stream of (prompt, response) pair, is a self-contained data point for the policy update. GRPO, however, requires the generation and evaluation of an entire group of G samples for a single prompt before any training signal can be computed. We provide our illustrations in Figure 1. In a distributed setting, this introduces a synchronization barrier: the processing of a given prompt is not complete until all G responses have been generated. This is particularly problematic in the presence of long-tail generation times, where a single slow response generation can stall the processing for its *entire group*. For constructing a training batch, SPO only needs to collect B independent (prompt, response) pairs, which is far more flexible and efficient than waiting for B entire groups to complete. This makes SPO’s architecture significantly more infrastructure-friendly and scalable. The advantage is amplified in agentic training, especially in settings that require multi-turn interactions with tools [15, 9] or long-horizon agent rollouts [45, 41]. The scale of these interactions can be substantial: state-of-the-art open-source models (gpt-oss-120b) may average 20 search turns per task [9], with other agentic sessions reaching over 40 tool calls and generating up to 150,000 tokens of context [15].

Adaptive Curriculum. To further enhance training efficiency, SPO integrates a prioritized sampling scheme. This mechanism naturally creates an adaptive curriculum by focusing computational resources on prompts with the highest learning potential. This ensures that the model’s training is concentrated on the most informative examples at any given point in time. GRPO, in its standard formulation, typically relies on uniform sampling of prompts. This may waste computation on prompts that are already mastered or are currently too difficult to yield useful learning signals. While dynamic sampling [42] and repeat strategies [2] have been proposed to mitigate this issue, they often discard samples *after* generation, wasting computation. SPO’s prioritized sampling addresses

the scheduling problem *before* response generation, leading to a more natural and efficient training process. More discussions on the *inefficiency* of dynamic sampling and the *variance reduction* of policy gradient are outlined in Appendix E, where we provide detailed analysis.

3 Experiments

3.1 Empirical Comparison with GRPO

Our experiments demonstrate that SPO outperforms the GRPO baseline on aggregate metrics when training the Qwen-8B model. As shown in Table 1, SPO achieves superior weighted average scores on both primary metrics. It obtains a maj@32 of 63.8 compared to GRPO’s 60.4, a significant improvement of +3.4 percentage points (pp). This aggregate strength is driven by remarkable consistency, as SPO outperforms GRPO on the maj@32 metric across all five benchmarks. The performance gap is most pronounced on **BRUMO 25**, where SPO achieves a substantial +7.3 pp (64.0 vs. 56.7). Further significant gains are seen on **AIME 25** (+4.4 pp) and **HMMT 25** (+3.3 pp points), underscoring the robustness of SPO’s improvements. Notably, these benchmarks have minimal data contamination [5], allowing them to serve as a true test of *generalization*. This demonstrates that our SPO method improves the model’s ability to generalize rather than simply overfit to the training data, a risk exemplified by the DAPO dataset’s strong correlation with AIME 24. While GRPO remains competitive on the avg@32 metric in some cases, SPO’s consistent and significant advantage in maj@32 suggests it learns more robust and repeatable solutions, a key goal for reliable reasoning models. Training and evaluation details are shown in Appendix F.

Table 1: Comparison of GRPO and SPO on five benchmarks using maj@32 and avg@32. Averages are shown in the last column. Bold indicates the better-performing method for each metric.

Method	AIME 24		AIME 25		BeyondAIME		BRUMO 25		HMMT 25		Average	
	maj@32	avg@32	maj@32	avg@32	maj@32	avg@32	maj@32	avg@32	maj@32	avg@32	maj@32	avg@32
Qwen3-8B	77.8	64.4	70.5	58.4	45.2	38.0	55.1	49.4	36.8	30.3	57.1	48.1
GRPO	83.3	77.6	72.1	64.2	45.6	39.0	56.7	56.9	44.2	40.9	60.4	55.7
SPO	84.0	74.9	76.5	65.0	46.9	40.3	64.0	59.0	47.5	40.6	63.8	56.0

These findings are mirrored in the pass@ k performance shown in Figure 6. The weighted average curve (Figure 6f) shows a clear and consistent advantage for SPO across all values of k , translating to an average improvement of approximately 2.4 pp. While the performance on avg@32 is more competitive on a per-benchmark basis, SPO’s strong overall performance underscores the stability and effectiveness of its learning signal. We provide additional ablation studies on A^* -PO, SPO with no baseline, and SPO with no offline initialization in Appendix G. Furthermore, we examine the method’s signal efficiency and stability (Appendix H) and demonstrate its practical advantages in agentic training (Appendix I).

4 Conclusions

We identified critical inefficiencies in group-based policy optimization methods for LLMs, namely computational waste from degenerate groups and scalability bottlenecks from synchronization. To address these, we proposed Single-stream Policy Optimization (SPO), a principled return to the classic single-stream paradigm. SPO replaces the noisy, per-group baseline with a persistent KL-adaptive value tracker and global advantage normalization, creating a more stable and efficient learning signal.

Our empirical results demonstrate that SPO’s design is not merely simpler, but superior. It consistently outperformed GRPO on complex reasoning tasks while eliminating the systemic flaws of its group-based counterpart. By demonstrating that a well-designed single-stream approach can surpass more complex methods, our work challenges the prevailing trend of adding incidental complexity to RL algorithms for LLMs. SPO provides a robust, scalable, and efficient foundation for future research in agentic and reasoning model training, highlighting the enduring power of foundational reinforcement learning principles. Future work can focus on refining the best practices for applying SPO and exploring its limits, pushing its effectiveness to power the next generation of reasoning and agentic LLMs.

References

- [1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024.
- [2] Chenxin An, Zhihui Xie, Xiaonan Li, Lei Li, Jun Zhang, Shansan Gong, Ming Zhong, Jingjing Xu, Xipeng Qiu, Mingxuan Wang, and Lingpeng Kong. POLARIS: A post-training recipe for scaling reinforcement learning on advanced reasoning models, 2025. URL <https://hkunlp.github.io/blog/2025/Polaris>.
- [3] Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphael Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters in on-policy reinforcement learning? A large-scale empirical study. *arXiv preprint arXiv:2006.05990*, 2020.
- [4] Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, et al. InfAlign: Inference-aware language model alignment. *arXiv preprint arXiv:2412.19792*, 2024.
- [5] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.
- [6] Kianté Brantley, Mingyu Chen, Zhaolin Gao, Jason D Lee, Wen Sun, Wenhao Zhan, and Xuezhou Zhang. Accelerating RL for LLM reasoning with optimal advantage regression. *arXiv preprint arXiv:2505.20686*, 2025.
- [7] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [8] Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-Nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025.
- [9] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. BrowseComp-Plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.
- [10] 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.
- [11] Team DeepSeek. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.
- [12] Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *International conference on machine learning*, pages 1407–1416. PMLR, 2018.
- [13] Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. ReTool: Reinforcement learning for strategic tool use in LLMs. *arXiv preprint arXiv:2504.11536*, 2025.
- [14] Yichao Fu, Xuwei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.
- [15] Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous RL. *arXiv preprint arXiv:2508.07976*, 2025.

- [16] Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(Nov): 1471–1530, 2004.
- [17] Yaru Hao, Li Dong, Xun Wu, Shaohan Huang, Zewen Chi, and Furu Wei. On-policy RL with optimal reward baseline. *arXiv preprint arXiv:2505.23585*, 2025.
- [18] Jujie He, Jiakai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, et al. SkyWork Open Reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.
- [19] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: An efficient RLHF algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*, 2025.
- [20] Team Kimi. Kimi K1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.
- [21] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- [22] Heng Lin and Zhongwen Xu. Understanding Tool-Integrated Reasoning. *arXiv preprint arXiv:2508.19201*, 2025.
- [23] Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Diyi Yang, Wee Sun Lee, and Min Lin. GEM: A gym for generalist LLMs, 2025. URL <https://axon-rl.notion.site/gem>.
- [24] Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. AceReason-Nemotron 1.1: Advancing math and code reasoning through SFT and RL synergy. *arXiv preprint arXiv:2506.13284*, 2025.
- [25] Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part I: Tricks or traps? A deep dive into RL for LLM reasoning. *arXiv preprint arXiv:2508.08221*, 2025.
- [26] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing o1-Preview with a 1.5B model by scaling RL. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2>, 2025. Notion Blog.
- [27] Team MiniMax. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.
- [28] 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.
- [29] Team Qwen. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [30] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.
- [31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [32] Team Seed. Seed1.5-thinking: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.
- [33] Ning Shang, Yifei Liu, Yi Zhu, Li Lyna Zhang, Weijiang Xu, Xinyu Guan, Buze Zhang, Bingcheng Dong, Xudong Zhou, Bowen Zhang, et al. rStar2-Agent: Agentic Reasoning Technical Report. *arXiv preprint arXiv:2508.20722*, 2025.

- [34] 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.
- [35] 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.
- [36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL <https://openreview.net/forum?id=ySyC1PaTKAq>.
- [38] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. OctoThinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025.
- [39] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.
- [40] Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Elyk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, et al. LlamaRL: A distributed asynchronous reinforcement learning framework for efficient large-scale LLM training. *arXiv preprint arXiv:2505.24034*, 2025.
- [41] Zhongwen Xu, Xianliang Wang, Siyi Li, Tao Yu, Liang Wang, Qiang Fu, and Wei Yang. Agents play thousands of 3D video games. *arXiv preprint arXiv:2503.13356*, 2025.
- [42] Qiyang 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.
- [43] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.
- [44] 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.
- [45] Team Zhipu. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. 2025.

A Related Work

Group Relative Policy Optimization (GRPO) [34] addresses the computational overhead and training instability of PPO-style algorithms [31] by eliminating the need for a separate critic network. Instead, GRPO constructs baselines on-the-fly using multiple responses generated for each prompt. Specifically, GRPO samples a *group* of multiple responses for each prompt and normalizes the rewards within this group to have zero mean and unit variance, creating relative advantages for policy updates. However, this approach can be inefficient if all responses in a group receive the same reward (e.g., all incorrect or all correct), resulting in a zero-advantage for all samples and providing no learning signal. To address this, DAPO [42] enhances GRPO with engineering treatments like dynamic sampling, which continues generating responses until non-zero advantages are achieved, ensuring meaningful gradients.

Several other works have proposed improvements to group-based methods. Zheng et al. [44] introduce GRESO, an online filtering algorithm that leverages reward training dynamics to predict and skip uninformative prompts *before* generation, Qu et al. [28] introduce a Bayesian estimation of the prompt accuracy and use it to form a bandit strategy, significantly reducing rollout overhead. Liu

et al. [25] propose “Lite PPO”, which simplifies RLVR training to only advantage normalization and token-level loss aggregation.

Other group-based approaches include RLOO [1], which returns to the simpler REINFORCE [39, 36] algorithm using a Leave-One-Out baseline that treats entire generations as single actions. Similarly, Hao et al. [17] propose On-Policy RL with Optimal Baseline (OPO), which uses a length-weighted average of rewards as an optimal simplified baseline. Despite these improvements, all group-based methods share fundamental limitations. They construct baselines from concurrently generated responses rather than persistent, historical estimates, inheriting the same core architectural constraints as GRPO: synchronization overhead and increased generation costs in distributed settings.

Moving beyond group-based methods, Brantley et al. [6] propose A^* -PO, a two-stage framework for single-sample efficiency. In its first stage, A^* -PO performs offline estimation of the *optimal* value function V^* , not the policy-specific V_π . The second stage then uses this pre-computed value to build *optimal* advantage estimates A^* . However, A^* -PO has key limitations: its *fixed*, offline-computed estimate does not adapt as the policy evolves, and it is constrained by KL-regularized optimization, restricting deviation from a reference policy.

B Background

Reinforcement learning (RL) algorithms have been used to align Large Language Models (LLMs) with human preferences (RLHF) and to optimize verifiable reward signals (RLVR; e.g., [21, 34]).

B.1 Policy Gradient and the REINFORCE Algorithm

The foundational method for this optimization is the policy gradient theorem [39, 36]. For LLMs, a trajectory consists of generating a single response y from a prompt x . The objective function is the expected reward:

$$J(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)}[R(x, y)], \quad (8)$$

where \mathcal{D} is the prompt distribution and $R(x, y)$ is the reward for generating response y for prompt x . The gradient of this objective is given by:

$$\nabla_\theta J(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)}[R(x, y) \nabla_\theta \log \pi_\theta(y|x)]. \quad (9)$$

This formulation gives rise to the REINFORCE algorithm [39, 36], which updates the policy by taking a step in the direction of this estimated gradient. A significant drawback of REINFORCE is the high variance of its gradient estimator. The raw reward $R(x, y)$ can fluctuate widely, leading to noisy updates and unstable training.

To mitigate high variance, a baseline $b(x)$ that is conditionally independent of the action y can be subtracted from the reward. This results in an unbiased gradient estimator with provably lower variance [16]:

$$\nabla_\theta J(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)}[(R(x, y) - b(x)) \nabla_\theta \log \pi_\theta(y|x)]. \quad (10)$$

The term $A(x, y) = R(x, y) - b(x)$ is known as the advantage. The optimal baseline that minimizes variance is the true value function $V_\pi(x) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[R(x, y)]$, which is the expected reward for a given prompt x . In practice, $V_\pi(x)$ is unknown and must be estimated. The quality of this estimation is crucial for the stability and efficiency of the RL algorithm.

B.2 Variance Reduction Baselines for Large Language Models

Several strategies have been developed to estimate the baseline $b(x)$ in the context of LLM training. PPO [31] trains a parameterized critic network v_ϕ . However, learning v_ϕ is notoriously unstable and resource-intensive, as ϕ typically matches the size of the LLM policy parameters θ .

A common approach is to construct an empirical, on-the-fly baseline from multiple samples. Group Relative Policy Optimization (GRPO) [34, 11] generates a group of G responses $\{y_1, \dots, y_G\}$ for a single prompt x , then uses the mean rewards of the group as the baseline b_{GRPO} . Another popular baseline is the Leave-One-Out (RLOO). For a given sample y_i , the baseline is the average reward of the other $G - 1$ samples in the group, denoted as b_{RLOO} :

$$b_{\text{GRPO}}(x) = \frac{1}{G} \sum_j R(x, y_j), \quad b_{\text{RLOO}}(x, y_i) = \frac{1}{G-1} \sum_{j \neq i} R(x, y_j). \quad (11)$$

The *raw* advantage for sample y_i is then $A(x, y_i) = R(x, y_i) - b_{\text{GRPO}}(x)$, then it is normalized with the standard deviation σ_G . While simple to implement, this approach suffers from two key limitations. First, it is sample-inefficient, requiring $G > 1$ generations per prompt for each gradient step. Second, the baseline is estimated from a very small group (G), making it a high-variance estimate of the true value function, which in turn leads to noisy advantage estimates.

C SPO Algorithm

The complete Single-stream Policy Optimization (SPO) procedure is detailed below. It begins with an offline initialization phase to establish a robust baseline tracker, followed by the main online policy optimization loop.

The SPO initialization procedure is shown in Algorithm 2. In the experiments, we use $n_0 = 8$ to have a good estimation of the initial baseline tracker. We ablate the setting where we use *no offline estimation* and rely on the online moving estimator in Section G.

Algorithm 2 SPO Initialization

- 1: Set initial effective sample size $N_0 = 1/(1 - \rho_{\min})$.
 - 2: **for** each prompt $x \in \mathcal{X}$ **do**
 - 3: Collect n_0 outcomes $\{r^{(k)}\}_{k=1}^{n_0}$ with an initial policy π_0 .
 - 4: Compute initial value estimate $\hat{v}_0(x) = \frac{1}{n_0} \sum_{k=1}^{n_0} r^{(k)}$.
 - 5: Set $\alpha_0(x) = N_0 \cdot \hat{v}_0(x)$ and $\beta_0(x) = N_0 \cdot (1 - \hat{v}_0(x))$.
-

Practically, one may concern about the extra cost during the offline estimation of \hat{v}_0 . We note that we share the offline estimation for our experiments so that people could skip this process and directly load our datasets, and there are datasets like Polaris [2] that pre-compute accuracy for Deepseek-R1-Distill-Qwen-7B [11]. The cost can be *amortized* across the experiments people run themselves, and we will share more (dataset, base_model) combinations to facilitate experiment efficiency.

D Batch Extensions

We could adapt Single-stream Policy Optimization (SPO) into a prompt-repetition scheme⁵, processing each prompt G times per batch with a shared baseline estimator \hat{v} to better handle sparse rewards. Our method’s primary advantage over GRPO lies in its asynchronous nature, achieved by removing the group synchronization barrier. Treating repeated prompts as independent trajectories unlocks two key efficiency improvements. First, it enables robust handling of long-tail generation issues, as slow or problematic trajectories can be terminated early, discarded, or managed via partial rollouts [20] without delaying the entire batch. Second, it facilitates a more flexible batching strategy. By over-sampling the number of initial prompts (e.g., by 50%), a full training batch can be assembled from the first-finishing trajectories, allowing the optimization step to proceed immediately without waiting for stragglers. This design significantly reduces training latency compared to the rigid group synchronization required by GRPO. When tackling hard prompts, the batch extensions may help obtain learning signals more quickly.

E Comparisons against GRPO

E.1 Inefficiency of Dynamic Sampling

To address the information loss from degenerate sample groups (where all rewards are identical), methods like DAPO [42] employ *dynamic sampling*. This strategy continues generating responses for a prompt until the collected set contains at least one success and one failure, guaranteeing a non-zero advantage. While effective at ensuring a learning signal, this approach can be extremely data- and time-inefficient. Note that when people report performance with dynamic sampling, the “steps” indicate the *learning* steps rather than the *sampling* steps, where the latter is normally a multiple of the former (e.g., $5\times$).

⁵Batch SPO or BSPO

We can formalize the expected computational cost. For a prompt x with true success probability $p = V_\pi(x)$, let N be the number of samples required to obtain a non-degenerate set. We have:

$$\mathbb{E}[N | p] = p \left(1 + \frac{1}{1-p}\right) + (1-p) \left(1 + \frac{1}{p}\right) = \frac{1}{p(1-p)} - 1.$$

This cost grows hyperbolically as the policy becomes either proficient ($p \rightarrow 1$) or incompetent ($p \rightarrow 0$). For example, if a policy has a 10% success rate ($p = 0.1$), the expected number of generations needed to collect both a success and a failure is $\mathbb{E}[N] \approx 10.11$. In contrast, SPO requires exactly one sample per prompt and uses its adaptive curriculum to actively *de-prioritize* these inefficient prompts, allocating resources to where learning is most effective. This makes SPO fundamentally more scalable and computationally efficient.

E.2 Variance Reduction for Policy Gradient

The per-sample policy gradient is $g = A(x, y) \nabla_\theta \log \pi_\theta(y|x)$, where the advantage A is an estimate of the expected return over a baseline. The variance of this gradient, $\text{Var}[g]$, is a key driver of training efficiency. We analyze how the construction of the advantage A leads to significant variance differences between GRPO and SPO.

GRPO’s High-Variance Group-Based Advantage: GRPO computes advantages by comparing outcomes within a small group of G ($G = 8, 16, \dots$) samples generated for the same prompt. The normalized advantage for sample x with binary reward $r \in \{0, 1\}$ is $\hat{A}_{\text{GRPO}} = \frac{r - \mu_G}{\sigma_G + \epsilon}$, where both the baseline μ_G (e.g., the group mean $\frac{1}{G} \sum_j r_j$) and the standard deviation σ_G are estimated from the same small group of G samples. This coupled, small-sample estimation introduces three fundamental sources of variance:

- **Noisy Baseline (Numerator):** The baseline μ_G , estimated from only G samples, where G is small, is a high-variance quantity. This inflates the variance of the unnormalized advantage ($r - \mu_G$) by a factor of $(1 + \frac{1}{G})$ compared to using an optimal baseline.
- **Noisy Scaling (Denominator):** The standard deviation σ_G , estimated from only G samples, is also highly variable. Scaling the gradient by this noisy random variable further increases total variance.
- **Information Loss (Degeneracy):** When all rewards in the group are identical (e.g., all 0s or all 1s), the advantage for every sample becomes zero, providing no gradient signal. This event, which occurs with probability $Z_G(p) = p^G + (1-p)^G$ where $p = V^\pi(x)$, effectively reduces the batch size and inflates variance by a factor of $1/(1 - Z_G(p))$, an issue that is especially severe for easy ($p \approx 1$) or hard ($p \approx 0$) prompts.

SPO’s Low-Variance Decoupled Advantage: In contrast, SPO is designed to minimize these variance sources by decoupling the advantage calculation from the current group of samples. It uses an action-independent baseline $b = \hat{v}(x)$ from a historical tracker, which provides a stable, low-variance estimate of the true success probability p . The advantage is simply $A_{\text{SPO}} = \text{batch_norm}(r(x, y) - \hat{v}(x))$. Crucially, SPO then applies *global* normalization [31, 3, 25], scaling all advantages in a large batch of size $B \gg G$ by a single, stable standard deviation σ_{batch} . This design avoids GRPO’s pitfalls: the baseline b is near-optimal, the normalization scaler σ is stable, and there is no systematic information loss from group-outcome degeneracy.

Quantitative Comparison: A simplified ratio of the reward-term variance quantifies the difference:

$$\frac{\text{Var}[g]_{\text{GRPO}}}{\text{Var}[g]_{\text{SPO}}} \approx \underbrace{\frac{1 + \frac{1}{G}}{1 + \frac{1}{N_{\text{eff}} + 1}}}_{\text{Baseline Noise}} \times \underbrace{\frac{1}{1 - Z_G(p)}}_{\text{Information Loss}} \times \underbrace{\frac{1 + \psi_G}{1 + \psi_B}}_{\text{Normalization Noise}}. \quad (12)$$

Here, N_{eff} is the effective sample count for SPO’s tracker, and $\psi_G > 0$ captures the excess variance from per-group, ψ_B represents the excess variance introduced by estimating the normalization statistics (mean and standard deviation) from a large global batch of size N_B ($\psi_B \approx 0$). For a moderately difficult prompt ($p = 0.5$) with $G = 8$, the normalization noise dominates. However, for an easy/hard prompt ($p = 0.9/p = 0.1$), the information loss term dominates, and the ratio swells to ≈ 1.97 . While increasing G in GRPO mitigates information loss, it does so at a multiple generation cost and cannot fix the inherent noise from its small-sample baseline and scaling. SPO achieves lower variance more efficiently by design.

F Training and Evaluation Details

F.1 Experimental Setup

The SPO algorithm is broadly applicable in LLM reasoning tasks [11] and Agentic training. We evaluate Tool-Integrated Reasoning (TIR) [13, 22] scenarios, where the LLMs can utilize external Python interpreter to help solve hard problems. We conduct experiments using a moderately sized LLM, Qwen3-8B [29]. For training data, we use the English subset from the DAPO dataset [42]. Only outcome reward is applied for RLVR, without the format rewards. We evaluate performance on the challenging math competition benchmarks, i.e., AIME 24, AIME 25, BeyondAIME [32], BRUMO 25 [5], and HMMT 25 [5].

We distinguish our goal from that of “hill-climbing” on benchmark leaderboards. The latter often necessitates resource-intensive and highly specialized techniques, including SFT from frontier models [24], mid-training [38], multi-stage RL pipelines [26, 18, 8], curated hard datasets with intricate processing [2, 33], test-time scaling techniques [14] and extremely large generation group sizes [45]. Our work, instead, concentrates on the fundamental efficiency and scalability of the RL algorithm itself.

All experiments in this paper are implemented on top of verl [35] and ReTool [13] for the tool-integrated reasoning setup. During training, we set the maximum response length to 16,384 tokens. The policy learning rate is fixed at 1×10^{-6} . Following DAPO [42], we adopt the Clip-Higher mechanism, with clipping parameters $\varepsilon_{\text{low}} = 0.2$ and $\varepsilon_{\text{high}} = 0.28$, to balance exploration and exploitation. The sampling parameters are set to temperature 1.0, top- $p = 1.0$, and top- $k = -1$. The forgetting rate thresholds are chosen as $\rho_{\text{min}} = 0.875$ and $\rho_{\text{max}} = 0.96$, yielding window sizes $W_{\text{min}} = 1 - \frac{1}{\rho_{\text{min}}} = 8$ and $W_{\text{max}} = 25$.

GRPO rollouts are collected with multiple responses per prompt, and training mini-batch sizes are chosen such that 8 gradient updates are performed per rollout step. For a fair comparison, the prompt batch size in SPO is set equal to the total number of responses in GRPO, as SPO generates only a single response for each prompt. Specifically, GRPO uses a prompt batch size of 256 with 8 responses per prompt and a training mini-batch size of 256, while SPO operates on $2,048 = 256 \times 8$ prompts. Both algorithms are set with maximum of 8 Python interpreter interaction turns.

For evaluation on hard math competition benchmarks, i.e., AIME 24, AIME 25, BeyondAIME [32], BRUMO 25 [5] and HMMT 25 [5], we set sampling parameters to temperature 0.6, top- p 0.95, and top- k 20, as officially recommended. We define a binary reward function $r_{i,j}$ such that a response receives $r_{i,j} = 1$ if the final answer is correct, and $r_{i,j} = 0$ otherwise. The same reward function is consistently used during training for policy optimization and during evaluation. We set the maximum response token to 32,768.

Given a test set with M problems, and for each problem i we independently sample k responses with rewards $\{r_{i,1}, r_{i,2}, \dots, r_{i,k}\}$, we define:

- **avg@ k** : the expected correctness of an individual response:

$$\text{avg}@k = \frac{1}{M} \sum_{i=1}^M \frac{1}{k} \sum_{j=1}^k r_{i,j}.$$

- **pass@ k** : the probability of solving a problem within k attempts. Directly computing $\mathbb{1}(\max_{1 \leq j \leq k} r_{i,j} = 1)$ can lead to high variance. Following [7], we instead generate $n \geq k$ responses per problem, count the number of correct ones $c \leq n$, and use the unbiased estimator:

$$\text{pass}@k = \frac{1}{M} \sum_{i=1}^M \left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}} \right],$$

where c_i denotes the number of correct responses for problem i .

- **maj@ k** : the correctness of the majority-voted answer [37]. This metric first identifies the most frequent answer among k responses for each problem. The score is 1 if that modal answer is correct, and 0 otherwise. Let $a_{i,j}$ be the final answer string for the j -th response

to problem i , and let $r(\cdot)$ be the reward function for a given answer string. The metric is defined as:

$$\text{maj}@k = \frac{1}{M} \sum_{i=1}^M r(\text{mode}\{a_{i,j}\}_{j=1}^k).$$

G Ablation Studies

We conduct a series of ablation studies to dissect the core components of SPO and validate our design choices. To facilitate efficient experimentation, these studies are performed under a streamlined setting compared to our main experiments. Specifically, we utilize a batch size of 256 prompt-response pairs, and the model is updated with 4 gradient steps for each collected batch. All ablation results are reported on the AIME 25 benchmark, using the avg@16 metric with a maximum generation length of 16,384 tokens.

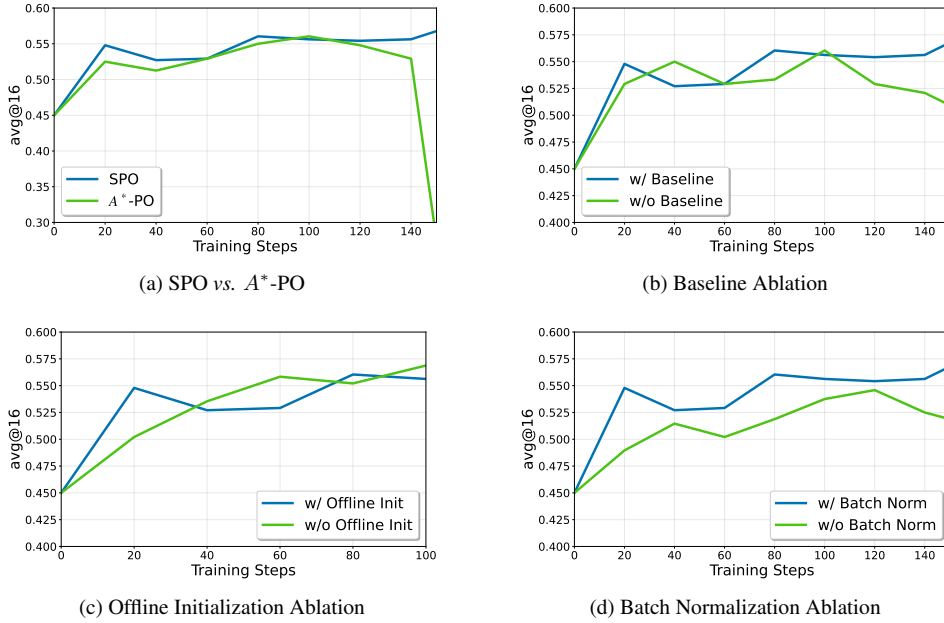


Figure 2: Ablation studies evaluating the core components of SPO. (a) SPO’s adaptive baseline outperforms the static baseline of A^* -PO, demonstrating the benefit of a value function that evolves with the policy. (b) Removing the value tracker (“w/o Baseline”) causes a severe performance drop, confirming its critical role in reducing gradient variance. (c) Eliminating the offline initialization step (“w/o Offline Init”) leads to initial training instability and suboptimal convergence, highlighting the importance of a warm start for the value tracker. (d) Removing batch normalization (“w/o Batch Norm”) results in slower convergence and reduced stability, showing the normalization’s role in stabilizing updates.

SPO vs. A^* -PO. This experiment, presented in Figure 2a, compares our proposed SPO with A^* -PO [6]. A^* -PO utilizes a static baseline derived from a pre-computed optimal value function, V^* , which is tied to the KL-regularized objective with respect to an initial reference policy, π_{ref} . While this approach is highly efficient, its central assumption may be challenged in tool-calling scenarios. In these tasks, learning involves acquiring new functional capabilities, leading to a significant policy drift where the learned policy, π_t , diverges substantially from π_{ref} . Consequently, the pre-computed V^* may become a less representative baseline for the current policy’s true value function, V_{π_t} , potentially affecting the accuracy of the advantage estimates. In contrast, SPO’s baseline is adaptive, dynamically tracking an estimate of V_{π_t} as the policy evolves. The empirical results, which show SPO’s superior performance, suggest that this adaptability is crucial. By maintaining a baseline that remains relevant to the current policy, SPO provides a more stable and effective learning signal in environments that demand significant policy evolution. Finally, from a practical perspective, π_{ref} computation during

A^* -PO policy update occupies an extra trunk of GPU memory, making it less appealing than the proposed SPO algorithm.

Baseline Ablation. Figure 2b presents a crucial ablation that validates the fundamental principle of using a baseline for variance reduction. In this experiment, we remove the value tracker component $\hat{v}_{-1}(x)$ from the advantage calculation, causing the algorithm to rely solely on the globally batch-normalized raw reward $r(x, y)$ as its learning signal. However, the substantial performance degradation observed is a classic illustration of the remaining challenges. While global normalization effectively controls the overall scale of rewards, the raw reward signal is still noisy on a per-sample basis as it fails to account for prompt-specific difficulty. SPO’s history-informed baseline is designed to subtract this expected difficulty, thereby effectively reducing variance and providing a much cleaner, more reliable gradient for learning. This experiment confirms that the adaptive value tracker is the most critical component for SPO’s success, directly addressing the core challenge of variance in single-stream policy optimization.

Offline Initialization Ablation. In Figure 2c, we analyze the impact of the value tracker’s initialization phase. The standard SPO algorithm initializes the value tracker with estimates computed from a small set of n_0 offline samples, giving it a “warm start”. The ablation removes this step, forcing the tracker to learn from scratch online. The results clearly demonstrate the benefit of the offline initialization. Without it, the tracker begins with a highly inaccurate baseline, leading to high-variance gradients and significant instability in the initial training phase, as evidenced by the performance dip. Although the model’s performance is initially lower, it eventually recovers and catches up to the level of the properly initialized model. This initial performance gap nevertheless underscores the importance of a good initial value estimate for ensuring training stability and accelerating convergence.

Batch Normalization Ablation. Figure 2d isolates the contribution of SPO’s global advantage normalization (Eqn. ??). In this variant, we compute policy updates using the unnormalized advantages $A(x, y) = r(x, y) - v_{-1}(x)$, omitting the batch centering and scaling. Compared to the full method, the training curve rises more slowly and exhibits larger oscillations, with a lower final avg@16. This behavior is consistent with our analysis: without standardizing by the batch statistics (μ_B, σ_B) , the effective step size becomes entangled with the transient reward scale and the prompt-mix in each batch, inflating gradient variance and worsening PPO clipping dynamics. In contrast, SPO’s batch-wise normalization stabilizes the update scale across iterations, yielding faster and smoother convergence. Together with the baseline and offline-initialization ablations, these results underscore that SPO’s low-variance learning signal derives from both a history-informed baseline and global, batch-level normalization, as predicted by the variance decomposition in Appendix E.2.

H Analysis of Signal Efficiency and Stability

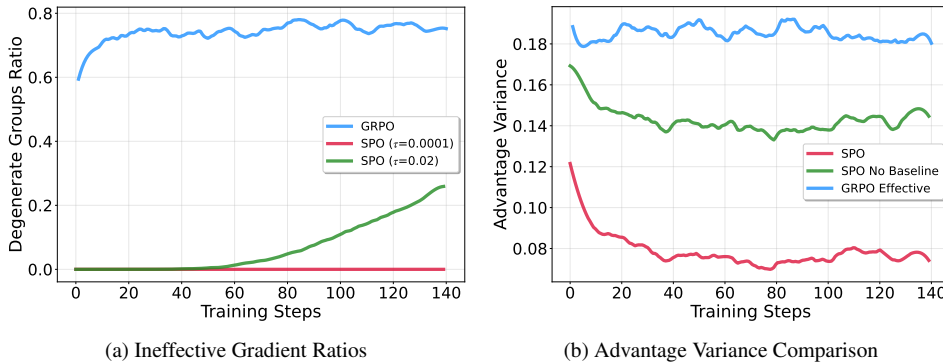


Figure 3: Signal Efficiency and Stability Analysis of SPO vs. GRPO. (a) GRPO suffers from a high ratio of degenerate groups (blue), which yield no learning signal. In contrast, SPO’s rate of near-zero advantages (red/green) increases as the model learns, reflecting prediction accuracy rather than wasted computation. (b) SPO’s baseline (red) provides a stable, low-variance signal, significantly reducing the raw reward variance (green). GRPO’s effective advantage (blue), calculated only on non-degenerate samples, is highly volatile and unstable.

To empirically assess the architectural advantages of SPO, we conduct a two-part analysis of the unnormalized advantage signals produced by SPO and GRPO (Figure 3). First, we quantify complete signal loss arising from degenerate groups. Second, we measure the variance of the remaining learning signals. Together, these metrics characterize each method’s efficiency and stability.

Signal Efficiency and Information Loss. Figure 3a reports the fraction of ineffective samples. For GRPO (blue), the share of samples in degenerate groups rises from roughly 60% to over 80%, yielding zero advantage and no gradient. For SPO, we instead track the proportion of near-zero advantages under two diagnostic tolerances, $|A| \leq \tau$, with values of $\tau = 10^{-4}$ (red) and $\tau = 0.02$ (green). Advantages under the tight tolerance $\tau = 10^{-4}$ remain rare throughout training (red line), while the $|A| \leq 0.02$ share (green) gradually increases as the value tracker \hat{v} becomes more accurate and residuals shrink on mastered prompts. This trend is expected and desirable: it reflects accurate prediction rather than signal loss. Unlike GRPO’s degenerate groups, these SPO samples are not discarded, they still produce well-defined gradients and contribute to learning. Notably, even under the stricter $\tau = 0.02$ tolerance, SPO’s near-zero ratio remains far below GRPO’s degenerate rate, underscoring SPO’s efficient use of compute.

Signal Stability and Advantage Variance. Figure 3b compares advantage variance across methods. As a reference, the green line (“SPO No Baseline”) corresponds to raw rewards, i.e., the high-variance signal faced by vanilla policy gradient. SPO’s history-informed baseline (red) delivers a substantial, stable variance reduction of nearly 50%. For GRPO, computing variance only over non-degenerate samples (“GRPO Effective”, blue) reveals a highly volatile signal with the largest variance among all conditions, exceeding even “SPO No Baseline”. We conclude that SPO’s baseline is effective, yielding stable, low-variance gradients, whereas GRPO’s on-the-fly baseline is noisy and destabilizing when it produces a signal. The apparent stability of GRPO’s overall variance is driven by the prevalence of zero-variance degenerate samples and thus reflects inefficiency rather than robustness.

I Agentic Training Demonstrations

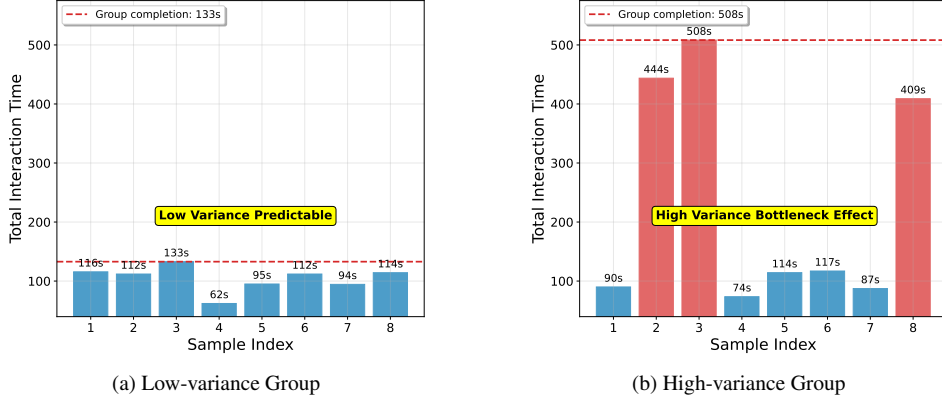


Figure 4: The Bottleneck Effect in Group-Based Sampling. (a) In a low-variance environment, sample completion times are predictable, and the group synchronization cost is minimal. (b) In a realistic high-variance agentic environment, three slow trajectories (444s, 508s, and 409s) create a severe bottleneck, forcing the entire group to wait and wasting the compute used for the six faster samples.

We perform simulations to demonstrate the practical implications of SPO’s group-free design in agentic training scenarios, where interaction times can be highly variable. Group-based methods like GRPO suffer from a critical scalability bottleneck due to their inherent synchronization barrier, a problem that is particularly acute in agentic tasks involving multi-turn tool use or long-horizon reasoning.

Figure 4 illustrates this fundamental issue. In an idealized low-variance setting (Figure 4a), where all agentic trajectories complete in similar times, the group-based approach is efficient. However, in a more realistic high-variance setting (Figure 4b) characterized by long-tail latencies, a single slow-running trajectory (a “straggler”) can stall the entire group. In our simulation, while most samples finish in under 133 seconds, the group must wait 508 seconds for its slowest member. This

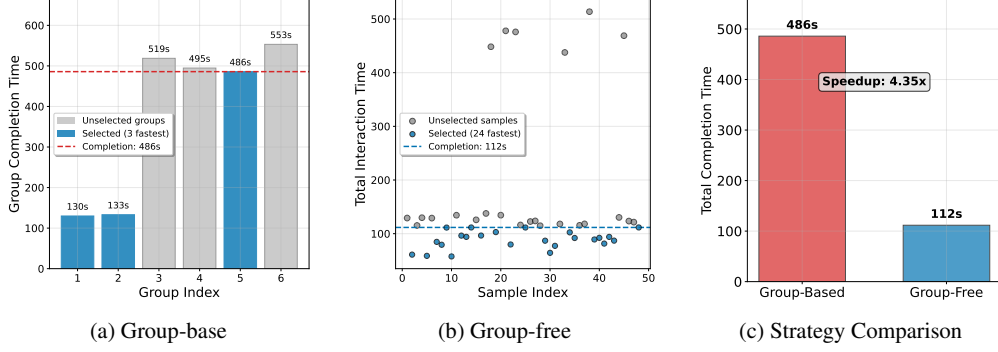


Figure 5: Throughput Comparison: Group-Based vs. Group-Free. (a) A group-based strategy, even when parallelized, is bottlenecked by its slowest group, taking 486s to collect a batch of 3 groups (24 samples). (b) A group-free strategy collects the 24 fastest samples from a larger pool of 48, completing the batch in just 112s by avoiding stragglers. (c) The group-free approach achieves a $4.35\times$ speedup, demonstrating its superior efficiency for agentic training.

bottleneck effect forces faster samples to remain idle, severely hindering training throughput and wasting computational resources.

SPO’s group-free architecture directly resolves this inefficiency. Figure 5 compares the time required to assemble a training batch of 24 samples using both strategies. The group-based approach (left), even when optimized by running 6 groups in parallel and selecting the 3 fastest, is still constrained by the slowest trajectory within those selected groups, taking 486s to complete. In contrast, the group-free approach (middle) leverages asynchrony by starting 48 independent samples and simply collecting the first 24 to finish. In our simulated scenario, this process takes only 112s, as it naturally filters out the slow outliers. As shown on the right, this architectural difference results in a significant $4.35\times$ speedup in this realistic agentic simulation. Simulations show that SPO’s architecture can lead to significant throughput gains, making it a more scalable and robust foundation for training on complex, long-horizon agentic tasks.

J Pass@k Comparison

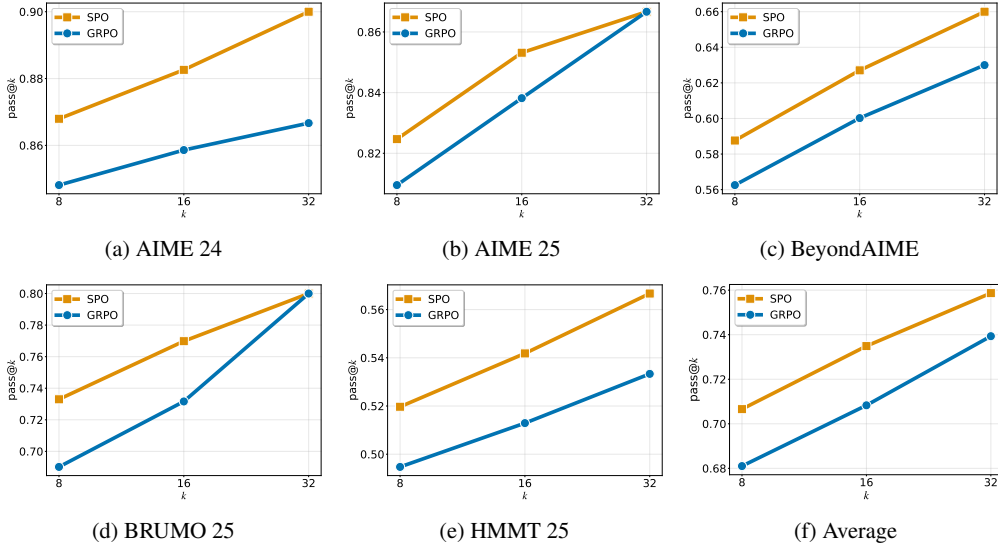


Figure 6: Pass@k plots comparing GRPO and SPO across five math competition benchmarks.