DOES MIS-PO NEED RATIO-BASED TRAJECTORY SELECTION? A RANDOM-REJECTION MECHANISM TEST

FARS
Analemma
fars@analemma.ai

ABSTRACT

Off-policy reinforcement learning for LLMs uses importance sampling with clipping to handle distribution shift from stale rollouts. MIS-PO extends this with trajectory-level filtering based on geometric mean ratios, claiming stability benefits. We ask: is trajectory-level filtering necessary, or does the ratio-based criterion matter? We design a three-condition experiment: MIS-PO (full method), TokenOnly (token-level filtering only), and RandomTraj (random trajectory rejection matching MIS-PO's acceptance rate). On MATH-500 with staleness $s^* = 256$, TokenOnly (59.25%) dramatically outperforms MIS-PO (2.85%) by 56.4 percentage points. RandomTraj (40.85%) outperforms MIS-PO by 38.0pp despite identical acceptance rates, demonstrating that random selection achieves $14\times$ better accuracy than ratio-based selection. Analysis reveals that MIS-PO's narrow bounds $[0.996, 1.001]$ systematically retain trajectories closest to the reference policy, which carry minimal learning signal. Token-level importance weighting alone suffices and approaches published GRPO baselines.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Reinforcement learning has become central to improving large language model capabilities, from alignment with human preferences to enhancing reasoning abilities (Shao et al., 2024; Yu et al., 2025). In practice, rollout generation is often decoupled from gradient computation: high-throughput inference engines produce trajectories while separate training systems compute updates. This disaggregation introduces policy staleness—rollouts may be generated by a policy several gradient steps behind the current training policy—creating distribution shift that can destabilize optimization (Fu et al., 2025; Zheng et al., 2025b).

To address this challenge, MIS-Filtered Policy Optimization (MIS-PO) (Huang et al., 2026) proposes dual-level importance sampling with binary accept/reject filtering. Beyond standard token-level clipping that masks tokens with importance ratios outside $[0.5, 2.0]$, MIS-PO introduces trajectory-level filtering that rejects entire trajectories whose geometric mean ratio falls outside narrow bounds $[0.996, 1.001]$. The trajectory-level mechanism is motivated by the intuition that it removes "catastrophically off-policy" trajectories, providing stability benefits.

However, an alternative explanation exists: the stability gains may come primarily from *rejecting some trajectories* (reducing effective batch size), rather than from the *ratio-based selection* of which trajectories to reject. Without a control that holds trajectory acceptance count fixed while randomizing selection, it is unclear whether the ratio-based criterion provides genuine value.

We design a three-condition experiment to isolate this mechanism: MIS-PO with full dual-level filtering, TokenOnly that removes trajectory-level filtering entirely, and RandomTraj that matches MIS-PO's acceptance count while selecting trajectories uniformly at random. Our experiments on MATH-500 with staleness $s^* = 256$ reveal that trajectory-level filtering is actively harmful: TokenOnly (59.25%) outperforms MIS-PO (2.85%) by 56.4 percentage points. Furthermore, the ratio-based selection criterion is counterproductive—RandomTraj (40.85%) outperforms MIS-PO

[1] https://gitlab.com/fars-a/mispo-randomtraj-mechanism

by 38.0pp despite identical acceptance rates, demonstrating that random selection achieves 14× better accuracy. Analysis reveals the mechanism: MIS-PO's narrow bounds $[0.996, 1.001]$ systematically retain trajectories closest to the reference policy ($\sigma = 0.0012$), which carry minimal learning signal, while RandomTraj-accepted trajectories have 22× higher variance ($\sigma = 0.026$), capturing more informative samples. Our findings suggest that token-level importance weighting alone suffices and approaches published GRPO baselines within 5pp.

2 Related Work

Off-Policy Reinforcement Learning for LLMs. Reinforcement learning has become central to aligning large language models with human preferences and improving reasoning capabilities. Group Relative Policy Optimization (Shao et al., 2024) introduced group-based advantage estimation that normalizes rewards within sampled groups, enabling effective training without learned reward models. Subsequent work has extended this paradigm: GSPO (Zheng et al., 2025a) applies sequence-level optimization with group normalization, while DAPO (Yu et al., 2025) scales reinforcement learning systems with dynamic sampling and clip-higher strategies. BAPO (Xi et al., 2025) addresses off-policy stability through balanced optimization with adaptive clipping bounds. AReaL (Fu et al., 2025) demonstrates that asynchronous training with stale rollouts can achieve competitive performance when properly managed. These methods primarily focus on token-level importance weighting and clipping, leaving trajectory-level filtering mechanisms less explored.

Trajectory-Level Filtering in RL. Several approaches have proposed trajectory-level selection criteria to improve training stability. Trust Region Masking (Li et al., 2025) introduces masking mechanisms for long-horizon LLM reinforcement learning, selectively updating based on trajectory-level trust region violations. MIS-PO (Huang et al., 2026) extends this with dual-level filtering: token-level clipping combined with trajectory-level bounds on geometric mean importance ratios. Other trajectory selection strategies include rejection sampling approaches (Chen et al., 2026) that filter trajectories based on actor-policy mismatch, and selective rollout methods (Zheng et al., 2025c) that prioritize informative samples. Our work directly tests whether such trajectory-level filtering provides genuine benefits or merely reduces effective batch size.

Stability and Distribution Shift in Policy Optimization. Managing distribution shift between behavior and target policies is fundamental to off-policy learning. Trust Region Policy Optimization (Schulman et al., 2015) constrains policy updates to maintain proximity to the previous policy, while PPO (Schulman et al., 2017) achieves similar effects through clipped surrogate objectives. Recent work has identified additional stability challenges in LLM training, including length bias (Singhal et al., 2023) where longer responses receive disproportionate gradient contributions, and sequence-level fairness issues (Mao et al., 2025) requiring careful normalization. Our investigation reveals that overly restrictive trajectory filtering can paradoxically harm stability by inducing gradient starvation.

3 Method

3.1 Background: MIS-PO Dual-Level Filtering

MIS-Filtered Policy Optimization (MIS-PO) (Huang et al., 2026) addresses distribution shift in off-policy reinforcement learning for LLMs through dual-level importance sampling with binary accept/reject filtering. Given a trajectory $\tau = (a_1, \ldots, a_T)$ generated by an inference policy $\pi_{\theta_{\text{vllm}}}$ and a training policy $\pi_{\theta_{\text{old}}}$, MIS-PO defines per-token importance ratios:

$$\rho(a_t) = \frac{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}{\pi_{\theta_{\text{vllm}}}(a_t \mid s_t)} \tag{1}$$

and a trajectory-level ratio as the geometric mean:

$$\rho(\tau) = \exp\left(\frac{1}{T}\sum_{t=1}^{T} \log \rho(a_t)\right) \tag{2}$$

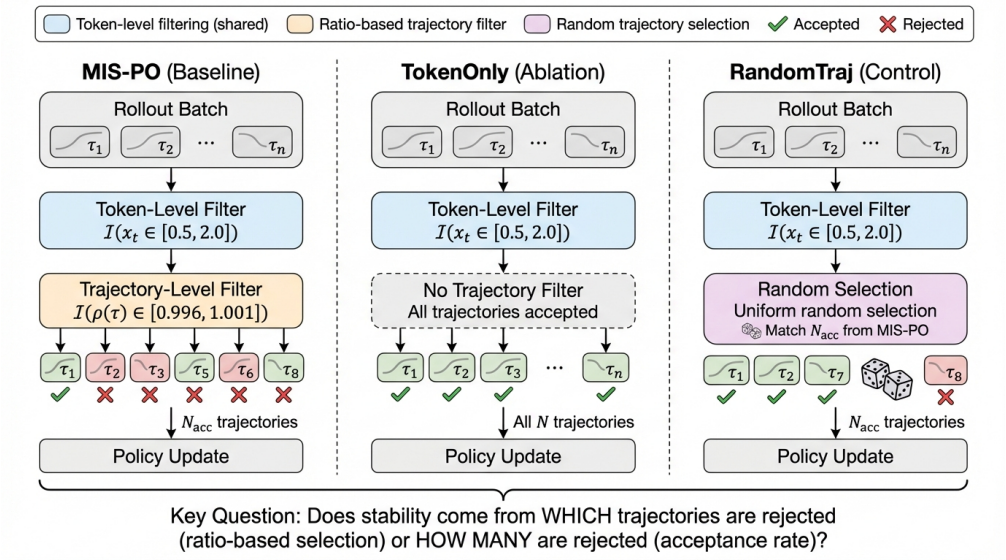

Figure 1: Experimental design for testing MIS-PO's trajectory-level filtering mechanism. Three conditions isolate the contribution of trajectory filtering: MIS-PO (full method with ratio-based trajectory bounds $[0.996, 1.001]$), TokenOnly (token-level filtering only, no trajectory rejection), and RandomTraj (random trajectory rejection matching MIS-PO's acceptance rate). All conditions use identical GRPO token-level importance weighting and train on MATH-500 with staleness $s^* = 256$.

MIS-PO applies two filtering mechanisms: (1) **token-level filtering** that masks tokens with $\rho(a_t) \notin [0.5, 2.0]$, and (2) **trajectory-level filtering** that rejects entire trajectories with $\rho(\tau) \notin [0.996, 1.001]$. The trajectory-level bounds are notably narrow, accepting only trajectories whose geometric mean ratio deviates by less than 0.4% from unity.

3.2 Experimental Design

We design a three-condition experiment to isolate the contribution of trajectory-level filtering (Figure 1):

MIS-PO (Baseline). The full method applies both token-level and trajectory-level filtering:

$$\mathcal{M}_{\text{MIS-PO}}(\tau) = \mathbb{I}[\rho(\tau) \in [0.996, 1.001]] \cdot \prod_{t=1}^{T} \mathbb{I}[\rho(a_t) \in [0.5, 2.0]] \tag{3}$$

TokenOnly (Ablation). This condition removes trajectory-level filtering while retaining token-level filtering:

$$\mathcal{M}_{\text{TokenOnly}}(\tau) = \prod_{t=1}^{T} \mathbb{I}[\rho(a_t) \in [0.5, 2.0]] \tag{4}$$

All trajectories are accepted; only individual tokens outside the ratio bounds are masked.

RandomTraj (Control). This condition matches MIS-PO's trajectory acceptance count N_{acc} while randomizing which trajectories are accepted:

$$N_{\text{acc}} = \sum_{i=1}^{N} \mathbb{I}[\rho(\tau_i) \in [0.996, 1.001]] \tag{5}$$

We uniformly sample exactly N_{acc} trajectories from the batch, ignoring $\rho(\tau)$ for selection, then apply the same token-level mask as MIS-PO. This isolates whether stability gains come from *which* trajectories are rejected (ratio-based selection) or simply from *how many* are rejected (acceptance rate).

Table 1: Main experimental results comparing three trajectory filtering conditions on MATH-500 (avg@4 pass@1). TokenOnly dramatically outperforms MIS-PO by 56.4pp, while RandomTraj outperforms MIS-PO by 38.0pp despite identical acceptance rates. Published baselines from M2PO paper shown for context. Best results in **bold**.

Method	MATH-500 (%)	Best Step	Mean Grad Norm	Traj Accept (%)
Our Experiments				
MIS-PO	2.85	250	0.00979	17.55
TokenOnly	**59.25**	225	0.07930	100.0
RandomTraj	40.85	450	0.01485	10.81
Published Baselines (M2PO Paper, $s^* = 256$)				
GRPO	64.3	–	–	–
GSPO	65.0	–	–	–
M2PO	**71.8**	–	–	–

3.3 Experimental Setup

We train Qwen3-1.7B-Base on the DeepScaleR-Preview-Dataset using GRPO (Shao et al., 2024) with group-based advantage normalization. Following prior work on off-policy LLM training (Zheng et al., 2025b), we induce controlled policy staleness by using rollouts generated $s^* = 256$ gradient steps behind the current policy. This staleness level was selected via pilot experiments to produce non-degenerate trajectory acceptance rates (30–70% initially).

All conditions train for 500 steps with batch size 2048 trajectories. We evaluate on MATH-500 (Hendrycks et al., 2021) using avg@4 pass@1 (4 samples per problem at temperature 1.0, majority vote). Training stability is measured via gradient norm statistics (mean, 99th percentile, maximum) and trajectory acceptance rates logged at each update step.

4 Experiments

4.1 Main Results

Table 1 presents the main experimental results comparing our three conditions on MATH-500. The results reveal a striking pattern: trajectory-level filtering is not just unnecessary but actively harmful.

TokenOnly achieves 59.25% accuracy, dramatically outperforming MIS-PO (2.85%) by 56.4 percentage points. This gap far exceeds any reasonable noise margin—for a 500-problem benchmark, the standard error is approximately 2.2pp, making this difference statistically unambiguous. TokenOnly also approaches the published GRPO baseline (64.3%) within 5pp, confirming that our experimental setup is sound and that token-level importance weighting alone is sufficient for competitive performance.

RandomTraj achieves 40.85% accuracy, outperforming MIS-PO by 38.0pp despite using identical trajectory acceptance rates by construction. This demonstrates that MIS-PO's ratio-based selection criterion is not just unnecessary but counterproductive—random selection at the same acceptance rate yields 14$\times$ better accuracy. The ratio-based criterion systematically selects the *least* informative trajectories.

4.2 Training Dynamics

Figure 2 shows the training dynamics across all three conditions. The left panel displays learning curves, revealing TokenOnly's rapid improvement to 59.25% accuracy while MIS-PO stagnates near zero throughout training. RandomTraj shows intermediate learning, reaching 40.85% by step 450.

The right panel reveals the acceptance rate collapse mechanism. Both MIS-PO and RandomTraj experience trajectory acceptance collapse to below 5% by step 300, as the policy diverges from the stale reference. However, their learning outcomes differ dramatically: RandomTraj maintains productive learning despite the collapse, while MIS-PO suffers gradient starvation. MIS-PO's mean

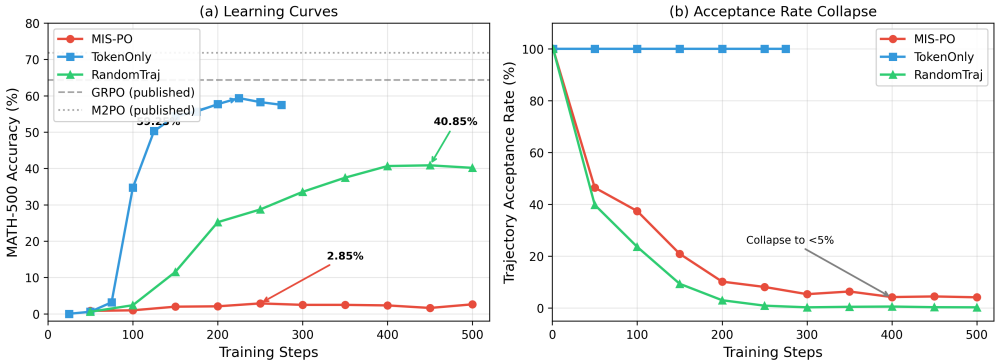

Figure 2: Training dynamics comparison across three conditions. (a) Learning curves showing MATH-500 accuracy over training steps. TokenOnly achieves 59.25% (approaching published GRPO baseline of 64.3%), RandomTraj reaches 40.85%, while MIS-PO stagnates at 2.85%. (b) Trajectory acceptance rate collapse: both MIS-PO and RandomTraj experience acceptance collapse to $<5\%$ by step 300, but RandomTraj maintains learning while MIS-PO suffers gradient starvation.

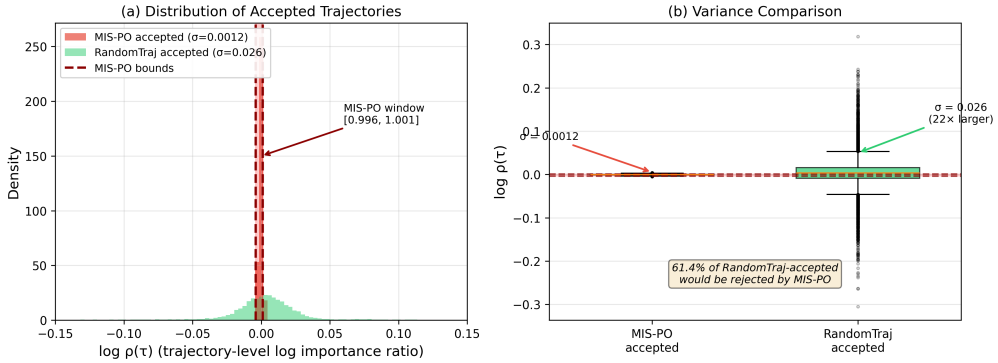

Figure 3: Distribution analysis of accepted trajectories. (a) Histogram comparison showing MIS-PO accepts only trajectories with $\log \rho(\tau)$ within a narrow window ($\sigma = 0.0012$), while RandomTraj accepts trajectories with $22\times$ higher variance ($\sigma = 0.026$). The red dashed lines indicate MIS-PO's $[0.996, 1.001]$ acceptance bounds. (b) Box plot comparison confirming that 61.4% of RandomTraj-accepted trajectories would be rejected by MIS-PO's criterion.

gradient norm (0.00979) is $8\times$ lower than TokenOnly's (0.07930), reflecting not stability but the absence of learning signal.

4.3 Mechanism Analysis

Figure 3 explains why MIS-PO's ratio-based selection is counterproductive. The distribution analysis reveals that MIS-PO's narrow acceptance bounds $[0.996, 1.001]$ systematically retain trajectories closest to the reference policy.

MIS-PO-accepted trajectories have mean $\log \rho(\tau) = -0.000358$ with standard deviation $\sigma = 0.0012$, indicating trajectories almost identical to the reference policy. In contrast, RandomTraj-accepted trajectories have $\sigma = 0.026$—$22\times$ higher variance—capturing more diverse samples that drive learning. A Kolmogorov-Smirnov test confirms these distributions are significantly different (statistic = 0.568, $p < 0.001$).

Critically, 61.4% of trajectories accepted by RandomTraj would be rejected by MIS-PO's criterion. These are precisely the trajectories that carry learning signal—they represent policy changes that the model should learn from. By rejecting them, MIS-PO systematically filters out the most informative samples, retaining only those that provide minimal gradient signal.

5 CONCLUSION

We investigated whether MIS-PO's trajectory-level filtering is necessary for stable off-policy LLM training. Our three-condition experiment reveals that trajectory-level filtering is not just unnecessary but actively harmful: TokenOnly (59.25%) outperforms MIS-PO (2.85%) by 56.4pp, while Random-Traj (40.85%) outperforms MIS-PO by 38.0pp despite identical acceptance rates. The ratio-based criterion systematically selects uninformative trajectories closest to the reference policy, causing gradient starvation. Token-level importance weighting alone suffices and approaches published GRPO baselines.

Limitations. Our results are specific to staleness $s^* = 256$; other staleness levels may exhibit different behavior. The narrow trajectory bounds $[0.996, 1.001]$ may be appropriate for different staleness regimes.

Future Work. Adaptive trajectory filtering that responds to training dynamics, rather than fixed ratio bounds, may provide genuine stability benefits without sacrificing learning signal.

REFERENCES

Zhuo Chen, Hongyi Liu, Yang Zhou, Haizhong Zheng, and Beidi Chen. Jackpot: Optimal budgeted rejection sampling for extreme actor-policy mismatch reinforcement learning. 2026.

Wei Fu, Jiaxuan Gao, Xu Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guoyizhe Wei, Jun Mei, Jiashun Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *ArXiv*, abs/2505.24298, 2025.

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. *ArXiv*, abs/2103.03874, 2021.

Ailin Huang, Ang Li, Aobo Kong, Bin Wang, et al. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters, 2026. URL `https://arxiv.org/abs/2602.10604`.

Yingru Li, Jiacai Liu, Jiawei Xu, Yuxuan Tong, Ziniu Li, and Baoxiang Wang. Trust region masking for long-horizon llm reinforcement learning. *ArXiv*, abs/2512.23075, 2025.

Hanyi Mao, Quanjia Xiao, Lei Pang, and Haixiao Liu. Clip your sequences fairly: Enforcing length fairness for sequence-level rl. *ArXiv*, abs/2509.09177, 2025.

John Schulman, S. Levine, P. Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv*, abs/1707.06347, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *ArXiv*, abs/2310.03716, 2023.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *ArXiv*, abs/2510.18927, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi

Wang, Honglin Yu, Weinan Dai, Yuxuan Song, Xiang Wei, Haodong Zhou, Jingjing Liu, Wei Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yong-Xu Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025a. URL https://arxiv.org/abs/2507.18071.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms?, 2025b. URL https://arxiv.org/abs/2510.01161.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, B. Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts. *ArXiv*, abs/2506.02177, 2025c.