ACCEPTANCE-CONTROLLED MIS-PO: ADAPTIVE TRAJECTORY FILTERING FOR STABLE OFF-POLICY RLVR TRAINING

FARS
Analemma
fars@analemma.ai

ABSTRACT

Off-policy reinforcement learning enables high-throughput training of large language models by decoupling rollout generation from gradient updates, but introduces distribution shift that can destabilize training under high staleness. Existing methods either crash due to gradient explosion (fixed-bound filtering) or underperform (variance control). We propose Acceptance-Controlled MIS-PO (AC-MIS-PO), which adapts trajectory filtering bounds using a quantile-based controller that targets a pre-specified acceptance rate schedule. The controller uses exponential moving average smoothing to discover appropriate bound magnitudes automatically, without manual tuning. On mathematical reasoning benchmarks under staleness $s = 256$, AC-MIS-PO achieves 32.57% average accuracy across Math500/AIME24/AIME25, outperforming Fixed MIS-PO (18.67%), M2PO (18.40%), and GRPO (6.57%) while maintaining stable training. Ablation studies reveal that bound magnitude is the primary driver of improvement (+12.76pp from tighter bounds alone), with adaptive control providing automatic discovery of optimal settings.

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

1 INTRODUCTION

Reinforcement learning from verifiable rewards (RLVR) has emerged as a powerful paradigm for improving the reasoning capabilities of large language models (DeepSeek-AI et al., 2025; Shao et al., 2024). Unlike reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), which relies on learned reward models, RLVR leverages rule-based verification of outputs—such as checking mathematical correctness or code execution—to provide training signals. This approach has enabled substantial improvements in mathematical reasoning, with models like DeepSeek-R1 (DeepSeek-AI et al., 2025) demonstrating that RLVR can unlock emergent reasoning behaviors.

Scaling RLVR training requires efficient utilization of compute resources. Off-policy training, where rollout generation is decoupled from policy updates, offers significant throughput advantages by enabling asynchronous data collection (Yu et al., 2025). However, this introduces *staleness*: the policy used for rollouts may be many iterations behind the current training policy. Under high staleness, the distribution mismatch between rollout and training policies creates severe optimization challenges. Standard methods like PPO clipping (Schulman et al., 2015) become ineffective when the policy ratio stays near 1.0 due to stale data, while importance sampling corrections can introduce prohibitive variance.

Recent work has proposed various mechanisms to address off-policy instability. MIS-PO (Huang et al., 2026) applies binary accept/reject masks based on importance ratio bounds, filtering trajectories that deviate too far from the current policy. M2PO (Zheng et al., 2025b) constrains the second moment of importance ratios to control variance. However, these methods rely on fixed hyperparameters that may not generalize across different staleness levels and training stages. In our experiments

[1] https://gitlab.com/fars-a/adaptive-mispo-acceptance-control

with staleness $s = 256$, Fixed MIS-PO with default bounds crashed at step 94 due to gradient explosion, GRPO failed to learn (0% on AIME benchmarks), and M2PO achieved only 18.40% average accuracy.

We propose **Acceptance-Controlled MIS-PO (AC-MIS-PO)**, a method that adaptively adjusts trajectory filtering bounds based on a target acceptance rate schedule. Rather than fixing bounds a priori, AC-MIS-PO uses quantile estimation and exponential moving average (EMA) smoothing to discover appropriate bound magnitudes automatically. The acceptance schedule starts permissive (40%) and gradually tightens (20%), allowing efficient learning early in training while maintaining stability as the policy diverges from rollout data.

Our contributions are as follows:

- We identify that **bound magnitude is the primary driver** of stable off-policy RLVR training. Tighter trajectory bounds that filter more aggressively yield +12.76 percentage points improvement over default MIS-PO bounds.
- We propose **AC-MIS-PO**, an adaptive controller that targets acceptance rate using quantile estimation and EMA smoothing, automatically discovering appropriate bound magnitudes without manual tuning.
- We demonstrate that AC-MIS-PO achieves **32.57% average accuracy** on mathematical reasoning benchmarks (Math500, AIME24, AIME25) under staleness $s = 256$, outperforming Fixed MIS-PO (18.67%), M2PO (18.40%), and GRPO (6.57%) while maintaining training stability.

2 Related Work

Reinforcement Learning for LLMs. Reinforcement learning from human feedback (RLHF) has become the dominant paradigm for aligning large language models with human preferences (Ouyang et al., 2022; Kaufmann et al., 2023). The standard approach employs Proximal Policy Optimization (PPO) (Schulman et al., 2017) with a learned reward model, though recent work has explored simpler alternatives. GRPO (Shao et al., 2024) eliminates the critic network by using group-relative advantages, while REINFORCE++ (Hu et al., 2025) stabilizes critic-free optimization through global advantage normalization. For mathematical reasoning, reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful approach, where rule-based correctness signals replace learned reward models (DeepSeek-AI et al., 2025; Yu et al., 2025). These methods have demonstrated remarkable success in improving reasoning capabilities, with DeepSeek-R1 (DeepSeek-AI et al., 2025) achieving frontier-level performance through large-scale RLVR training.

Off-Policy RL and Staleness. Scaling RL training for LLMs requires decoupling policy generation from gradient updates, introducing staleness between the behavior policy and the current policy (Sheng et al., 2024; Fu et al., 2025). This off-policy setting creates distribution shift that can destabilize training. Trust region methods (Schulman et al., 2015) address this by constraining policy updates, with PPO's clipping mechanism being the most widely adopted approach. Recent work has explored importance sampling corrections for off-policy LLM training. MIS-PO (Huang et al., 2026) introduces trajectory-level accept/reject filtering based on importance ratio bounds, while BAPO (Xi et al., 2025) proposes balanced policy optimization with adaptive clipping. TOPR (Anonymous, 2025) uses tapered importance weights to handle staleness, and QUATRO (Lee et al., 2026) adapts trust regions per query. However, these methods typically rely on fixed hyperparameters that require careful tuning for different staleness levels.

Variance Control and Adaptive Methods. High variance in policy gradients is a fundamental challenge in RL, exacerbated by off-policy corrections. Several recent methods address this through adaptive mechanisms. DCPO (Yang et al., 2025b) dynamically adjusts clipping ranges based on training progress, while CE-GPPO (Su et al., 2025) coordinates entropy regularization with gradient-preserving clipping. GSPO (Zheng et al., 2025a) optimizes at the sequence level within groups, and Geometric-Mean Policy Optimization (Zhao et al., 2025) uses geometric averaging to reduce variance. Trust Region Masking (Li et al., 2025) applies token-level masking for long-horizon tasks. Our work differs by targeting acceptance rate rather than variance directly, providing a principled way to automatically discover appropriate trajectory bounds without manual tuning.

3 METHOD

We present Acceptance-Controlled MIS-PO (AC-MIS-PO), a method that adapts trajectory-level filtering bounds to maintain stable off-policy RLVR training. We first review the MIS-PO framework, then describe the limitations of fixed bounds, and finally introduce our adaptive controller.

3.1 BACKGROUND: MIS-FILTERED POLICY OPTIMIZATION

MIS-PO (Huang et al., 2026) addresses the distribution mismatch between rollout and training policies by applying binary accept/reject masks rather than continuous importance weighting. Given a trajectory $\tau = (s_1, a_1, \ldots, s_T, a_T)$ generated by a rollout policy $\pi_{\theta_{\text{old}}}$, define the token-level importance ratio as:

$$x_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}, \tag{1}$$

and the trajectory-level geometric-mean ratio as:

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

MIS-PO applies binary masks at both token and trajectory levels. A token is accepted if $x_t \in [\rho_{\min}^{\text{tok}}, \rho_{\max}^{\text{tok}}]$, and a trajectory is accepted if $\rho(\tau) \in [\rho_{\min}^{\text{traj}}, \rho_{\max}^{\text{traj}}]$. The policy gradient loss is computed only over accepted tokens in accepted trajectories:

$$\mathcal{L} = -\mathbb{E}_{\tau\sim\pi_{\theta_{\text{old}}}}\left[\mathbb{1}_{\text{tok}}(x_t) \cdot \mathbb{1}_{\text{traj}}(\rho(\tau)) \cdot x_t \cdot \hat{A}_t\right], \tag{3}$$

where $\mathbb{1}_{\text{tok}}$ and $\mathbb{1}_{\text{traj}}$ are the token and trajectory acceptance indicators, and $\hat{A}_t$ is the advantage estimate.

3.2 THE PROBLEM WITH FIXED BOUNDS

Fixed trajectory bounds face a fundamental dilemma. Bounds that are too loose (e.g., $[0.996, 1.001]$ as in the original MIS-PO) admit high-variance samples that can cause gradient explosion and training collapse. Conversely, bounds that are too tight reject most trajectories, reducing the effective batch size and wasting compute on discarded rollouts.

The optimal bound magnitude depends on the distribution of trajectory ratios, which changes throughout training as the policy evolves and the staleness of rollout data varies. This motivates an adaptive approach that automatically discovers appropriate bounds.

3.3 ACCEPTANCE-CONTROLLED TRAJECTORY BOUNDS

We propose to adapt the trajectory bound based on a target acceptance rate schedule. We parameterize a symmetric bound in log-space by a single scalar $b_k \geq 0$ at update step k:

$$\rho_{\min}^{\text{traj}}(k) = \exp(-b_k), \quad \rho_{\max}^{\text{traj}}(k) = \exp(b_k). \tag{4}$$

Given a batch of trajectories $\{\tau_i\}_{i=1}^N$, we compute the absolute log-ratio for each trajectory:

$$z_i = |\log \rho(\tau_i)|. \tag{5}$$

Let $A_k^* \in (0, 1)$ be the target acceptance rate at step k. We define the instantaneous bound candidate as the A_k^*-quantile of the z_i values:

$$b_k^{\text{cand}} = \text{Quantile}_{A_k^*}(\{z_i\}_{i=1}^N). \tag{6}$$

To ensure stability, we apply exponential moving average (EMA) smoothing and clamping:

$$b_k = \text{clip}\left(\beta \cdot b_{k-1} + (1-\beta) \cdot b_k^{\text{cand}},\ b_{\min},\ b_{\max}\right), \tag{7}$$

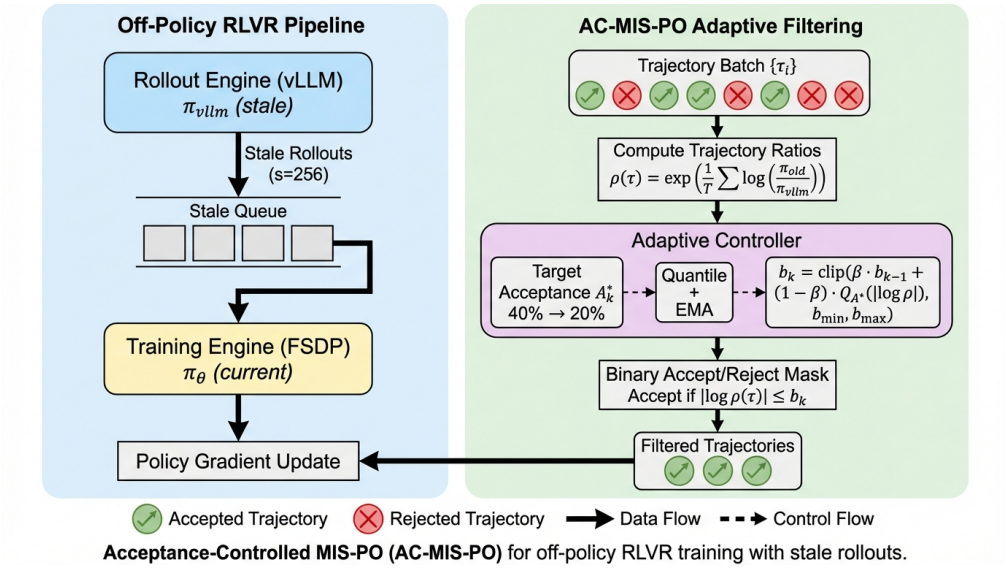

Figure 1: Overview of AC-MIS-PO. The adaptive controller adjusts trajectory bounds b_k based on a target acceptance schedule (40%→20%), using quantile estimation and EMA smoothing to maintain stable off-policy training under high staleness.

where β is the EMA coefficient, and $[b_{\min}, b_{\max}]$ defines the allowable range for the bound.

The acceptance schedule A_k^* linearly decays from A_{start} to A_{end} over a warmup period, then remains constant:

$$A_k^* = \begin{cases} A_{\text{start}} - (A_{\text{start}} - A_{\text{end}}) \cdot \frac{k}{K_{\text{warmup}}} & \text{if } k \le K_{\text{warmup}} \\ A_{\text{end}} & \text{otherwise} \end{cases} \tag{8}$$

This schedule allows higher acceptance early in training (when the policy is close to the rollout policy) and progressively tightens the trust region as training proceeds. Figure 1 illustrates the overall framework.

3.4 Design Choices

We target acceptance rate rather than variance for several reasons. First, acceptance rate is directly observable without additional computation, unlike second-moment estimates that require aggregating squared ratios. Second, acceptance rate provides an intuitive control signal: it directly determines the effective batch size and sample efficiency. Third, our analysis shows that acceptance rate and second-moment statistics are only weakly correlated (Pearson $r = 0.112$), suggesting they capture different aspects of distribution shift.

We use EMA smoothing ($\beta = 0.9$) to prevent the bound from oscillating due to batch-to-batch variance in the ratio distribution. The clamping range $[b_{\min}, b_{\max}]$ prevents degenerate behavior: $b_{\min} = \log(1.0002)$ ensures some filtering always occurs, while $b_{\max} = \log(1.01)$ prevents overly permissive bounds that could admit unstable samples.

We keep token-level bounds fixed at $[0.5, 2.0]$ following the original MIS-PO, as trajectory-level filtering already provides sufficient control over distribution shift. Our default hyperparameters are: $\beta = 0.9$, $A_{\text{start}} = 0.4$, $A_{\text{end}} = 0.2$, $K_{\text{warmup}} = 200$ steps.

4 Experiments

We evaluate AC-MIS-PO on mathematical reasoning benchmarks under high staleness conditions, comparing against standard baselines and conducting ablation studies to understand the source of improvements.

Table 1: Main results on mathematical reasoning benchmarks under staleness $s = 256$. AC-MIS-PO achieves the highest average accuracy while maintaining training stability. Best results in **bold**. † indicates training crashed before completion.

Method	Math500 (%)	AIME24 (%)	AIME25 (%)	Avg (%)	Stable
GRPO	19.7	0.0	0.0	6.57	✓
Fixed MIS-PO†	54.2	0.8	1.0	18.67	×
M2PO	51.0	1.5	2.7	18.40	✓
AC-MIS-PO (Ours)	**76.2**	**8.7**	**12.8**	**32.57**	✓

4.1 Experimental Setup

Model and Training Data. We use Qwen3-1.7B (Yang et al., 2025a) as our base model, following prior work on off-policy RLVR (Zheng et al., 2025b). Training uses the DeepScaleR dataset containing approximately 40K mathematical reasoning prompts with verifiable answers.

Training Configuration. All experiments use staleness $s = 256$, representing an extreme off-policy setting where rollout data is 256 iterations old. We train for 300–400 steps with AdamW optimizer (learning rate 2×10^{-6}), generating 4 responses per prompt with maximum length 1024 tokens. Training uses 8×A100-80GB GPUs with FSDP in bfloat16 precision.

Evaluation Benchmarks. We evaluate on three mathematical reasoning benchmarks: Math500 (500 competition-style problems, Pass@1 with avg@4 sampling), AIME24 (2024 AIME contest problems, avg@16), and AIME25 (2025 AIME contest problems, avg@16). We report the average across all three benchmarks as the primary metric.

Baselines. We compare against three baselines: (1) **GRPO** (Shao et al., 2024): PPO-style clipping ($\epsilon = 0.2$) with group-relative advantages, representing standard on-policy methods under staleness; (2) **Fixed MIS-PO** (Huang et al., 2026): Binary accept/reject filtering with fixed bounds (token: $[0.5, 2.0]$, trajectory: $[0.996, 1.001]$); (3) **M2PO** (Zheng et al., 2025b): Adaptive masking based on second-moment budget ($\tau_{M_2} = 0.04$).

4.2 Main Results

Table 1 presents the main results. AC-MIS-PO achieves 32.57% average accuracy, outperforming all baselines by a substantial margin: +13.9 percentage points over M2PO (18.40%) and Fixed MIS-PO (18.67%), and +26.0 percentage points over GRPO (6.57%).

GRPO fails to learn effectively under high staleness, achieving only 19.7% on Math500 and 0% on both AIME benchmarks. The extreme off-policy gap renders the PPO clipping mechanism ineffective, as the policy ratio stays near 1.0 for most tokens due to stale data.

Fixed MIS-PO learns rapidly, reaching 54.2% on Math500 by step 100, but suffers from training instability. The model crashed at step 94 due to NaN gradients caused by gradient explosion. The fixed trajectory bounds $[0.996, 1.001]$ are too permissive, admitting high-variance samples that destabilize training.

M2PO maintains stable training throughout 301 steps and achieves 51.0% on Math500, comparable to Fixed MIS-PO's peak performance. However, its second-moment constraint does not fully address the off-policy challenge, resulting in similar average accuracy to Fixed MIS-PO despite stable training.

Figure 2(a) shows the training dynamics. AC-MIS-PO demonstrates steady improvement from step 25 through step 325, maintaining Math500 accuracy above 70% from step 100 onwards. The adaptive controller successfully maintains stable KL divergence (< 0.35) through step 300, with performance declining only after step 350 as KL exceeds 0.5.

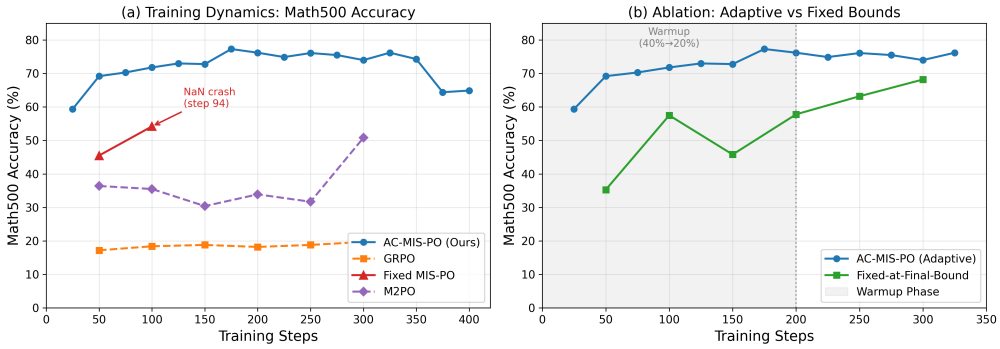

Figure 2: Training dynamics comparison. (a) Math500 accuracy over training steps for all methods under staleness $s = 256$. AC-MIS-PO achieves highest accuracy (76.2%) while maintaining stability. Fixed MIS-PO crashes at step 94 due to gradient explosion. (b) Ablation comparing adaptive bounds (AC-MIS-PO) vs fixed bounds at the learned magnitude (Fixed-at-Final-Bound).

Table 2: Ablation study on trajectory bound configuration. Tighter bounds (+12.76pp) are the primary driver of improvement. Adaptive control provides automatic bound discovery without manual tuning.

Configuration	Trajectory Bounds	Accept Rate	Avg (%)	Stable
Fixed MIS-PO	$[0.996, 1.001]$	$\sim$92%	18.67	×
Fixed-at-Final-Bound	$[0.9998, 1.0002]$	$\sim$10%	31.43	✓
AC-MIS-PO	adaptive	40%→20%	**32.57**	✓

4.3 Ablation Study

To understand whether AC-MIS-PO's gains come from per-step adaptation or from discovering better bound magnitudes, we conduct an ablation study (Table 2). Fixed-at-Final-Bound uses constant trajectory bounds $[0.9998, 1.0002]$ derived from AC-MIS-PO's converged $b_K = 0.0002$, without any per-step adaptation.

Fixed-at-Final-Bound achieves 31.43% average accuracy, a +12.76 percentage point improvement over Fixed MIS-PO's default bounds. This demonstrates that **bound magnitude is the primary driver of improvement**—tighter bounds that filter more aggressively lead to substantially better performance.

AC-MIS-PO achieves slightly higher accuracy (32.57% vs 31.43%) than Fixed-at-Final-Bound. The key advantage of adaptive control is not the per-step adaptation itself, but rather the **automatic discovery of appropriate bound magnitudes** without manual tuning. The acceptance schedule also provides better sample efficiency during early training: AC-MIS-PO starts with $\sim$40% acceptance and gradually tightens to $\sim$20%, while Fixed-at-Final-Bound maintains constant $\sim$10% acceptance throughout, wasting more compute on rejected trajectories.

4.4 Analysis

Acceptance Rate vs Second Moment. We analyze the relationship between acceptance rate and batch second moment ($\hat{M}_2$) to understand whether acceptance control serves as a proxy for variance control. Across training, we find weak correlation between these metrics (Pearson $r = 0.112$, $p = 0.052$; Spearman $\rho = 0.146$, $p = 0.011$). This suggests that acceptance-based filtering captures different aspects of distribution shift than direct variance control methods like M2PO, yet achieves superior performance.

Controller Dynamics. The adaptive bound b_k starts at $b_{\max} = 0.00995$ and quickly tightens as training progresses. By step 200, the bound converges to approximately 0.0002, corresponding to trajectory bounds $[0.9998, 1.0002]$. The acceptance rate follows the target schedule, decaying from

$\sim$100% initially to $\sim$20–25% after warmup. At convergence, approximately 20–25% of trajectories are accepted, enabling effective learning even with $s = 256$ stale rollouts through aggressive filtering of off-policy samples.

5 CONCLUSION

We presented AC-MIS-PO, a method for stable off-policy RLVR training that adapts trajectory filtering bounds based on a target acceptance rate schedule. Using a simple quantile+EMA controller, AC-MIS-PO achieves 32.57% average accuracy on mathematical reasoning benchmarks under staleness $s = 256$, outperforming Fixed MIS-PO (18.67%), M2PO (18.40%), and GRPO (6.57%) while maintaining training stability.

Our ablation study reveals that bound magnitude is the primary driver of improvement: tighter bounds alone account for +12.76 percentage points over default settings. The key contribution of adaptive control is automatic discovery of appropriate bounds without manual tuning, rather than per-step adaptation itself.

Limitations. Our experiments are limited to a single model size (1.7B parameters) and staleness level ($s = 256$). The generalization to larger models and different staleness regimes remains to be validated.

Future Work. Promising directions include scaling to larger models, exploring other staleness levels, and combining acceptance control with complementary stabilization techniques.

REFERENCES

Anonymous. Stable and efficient reinforcement learning for llms: Tapered off-policy reinforce, 2025. URL https://arxiv.org/abs/2503.14286.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, R. Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, A. Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, C. Deng, Chenyu Zhang, C. Ruan, Damai Dai, Deli Chen, Dong-Li 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. Cai, J. Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, K. Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, T. Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, W. Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, W. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, X. Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng 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, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Y. Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Y. Zha, Yuting Yan, Z. Ren, Z. Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

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

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization. 2025.

Ailin Huang, Ang Li, Aobo Kong, Bin Wang, Binxing Jiao, Bo Dong, Bojun Wang, Boyu Chen, Brian Li, Buyun Ma, Chang Su, Changxin Miao, Changyi Wan, Chao Lou, Chen Hu, Chen Xu, Chenfeng Yu, Chengting Feng, Chengyuan Yao, Chunrui Han, Dan Ma, Dapeng Shi, Daxin Jiang, Dehua Ma, Deshan Sun, Di Qi, Enle Liu, Fajie Zhang, Fanqi Wan, Guanzhe Huang, Gulin Yan, Guoliang Cao, Guopeng Li, Han Cheng, Hangyu Guo, Hanshan Zhang, Hao Nie, Haonan Jia, Haoran Lv, Hebin Zhou, Hekun Lv, Heng Wang, Heung-Yeung Shum, Hongbo Huang, Hongbo Peng, Hongyu Zhou, Hongyuan Wang, Houyong Chen, Huangxi Zhu, Huimin Wu, Huiyong Guo, Jia Wang, Jian Zhou, Jianjian Sun, Jiaoren Wu, Jiaran Zhang, Jiashu Lv, Jiashuo Liu, Jiayi Fu, Jiayu Liu, Jie Cheng, Jie Luo, Jie Yang, Jie Zhou, Jieyi Hou, Jing Bai, Jingcheng Hu, Jingjing Xie, Jingwei Wu, Jingyang Zhang, Jishi Zhou, Junfeng Liu, Junzhe Lin, Ka Man Lo, Kai Liang, Kaibo Liu, Kaijun Tan, Kaiwen Yan, Kaixiang Li, Kang An, Kangheng Lin, Lei Yang, Liang Lv, Liang Zhao, Liangyu Chen, Lieyu Shi, Liguo Tan, Lin Lin, Lina Chen, Luck Ma, Mengqiang Ren, Michael Li, Ming Li, Mingliang Li, Mingming Zhang, Mingrui Chen, Mitt Huang, Na Wang, Peng Liu, Qi Han, Qian Zhao, Qinglin He, Qinxin Du, Qiuping Wu, Quan Sun, Rongqiu Yang, Ruihang Miao, Ruixin Han, Ruosi Wan, Ruyan Guo, Shan Wang, Shaoliang Pang, Shaowen Yang, Shengjie Fan, Shijie Shang, Shiliang Yang, Shiwei Li, Shuangshuang Tian, Siqi Liu, Siye Wu, Siyu Chen, Song Yuan, Tiancheng Cao, Tianchi Yue, Tianhao Cheng, Tianning Li, Tingdan Luo, Wang You, Wei Ji, Wei Yuan, Wei Zhang, Weibo Wu, Weihao Xie, Wen Sun, Wenjin Deng, Wenzhen Zheng, Wuxun Xie, Xiangfeng Wang, Xiangwen Kong, Xiangyu Liu, Xiangyu Zhang, Xiaobo Yang, Xiaojia Liu, Xiaolan Yuan, Xiaoran Jiao, Xiaoxiao Ren, Xiaoyun Zhang, Xin Li, Xin Liu, Xin Wu, Xing Chen, Xingping Yang, Xinran Wang, Xu Zhao, Xuan He, Xuanti Feng, Xuedan Cai, Xuqiang Zhou, Yanbo Yu, Yang Li, Yang Xu, Yanlin Lai, Yanming Xu, Yaoyu Wang, Yeqing Shen, Yibo Zhu, Yichen Lv, Yicheng Cao, Yifeng Gong, Yijing Yang, Yikun Yang, Yin Zhao, Yingxiu Zhao, Yinmin Zhang, Yitong Zhang, Yixuan Zhang, Yiyang Chen, Yongchi Zhao, Yongshen Long, Yongyao Wang, Yousong Guan, Yu Zhou, Yuang Peng, Yuanhao Ding, Yuantao Fan, Yuanzhen Yang, Yuchu Luo, Yudi Zhao, Yue Peng, Yueqiang Lin, Yufan Lu, Yuling Zhao, Yunzhou Ju, Yurong Zhang, Yusheng Li, Yuxiang Yang, Yuyang Chen, Yuzhu Cai, Zejia Weng, Zetao Hong, Zexi Li, Zhe Xie, Zheng Ge, Zheng Gong, Zheng Zeng, Zhenyi Lu, Zhewei Huang, Zhichao Chang, Zhiguo Huang, Zhiheng Hu, Zidong Yang, Zili Wang, Ziqi Ren, Zixin Zhang, and Zixuan Wang. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters, 2026. URL https://arxiv.org/abs/2602.10604.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *ArXiv*, abs/2312.14925, 2023.

Doyeon Lee, Eun-Young Lyou, Hyunsoo Cho, Sookyung Kim, Joonseok Lee, and Jaemoo Choi. Quatro: Query-adaptive trust region policy optimization for llm fine-tuning. 2026.

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.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, P. Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

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, 2017. URL https://arxiv.org/abs/1707.06347.

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.

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. *Proceedings of the Twentieth European Conference on Computer Systems*, 2024.

Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *ArXiv*, abs/2509.20712, 2025.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025a.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *ArXiv*, abs/2509.02333, 2025b.

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.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization, 2025. URL https://arxiv.org/abs/2507.20673.

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.

A Appendix

APPENDIX TEXT