PREFIX-RATIO GRPO: IMPROVING GRADIENT QUALITY FOR REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Distributed reinforcement learning (RL) systems for large language models (LLMs) decouple rollout generation from learning, introducing staleness where trajectories are generated by behavior policies that lag behind the current learner. Standard importance sampling corrections apply per-token ratios that treat each token independently, ignoring sequential dependencies in autoregressive generation. We propose Prefix-Ratio GRPO, which incorporates prefix information into importance ratios: if any prefix token has become unlikely under the current policy, all subsequent tokens are downweighted. Our prefix-aware ratio $\tilde{\rho}_t = \underline{\rho}_t \cdot \rho_t$, where $\underline{\rho}_t = \min_{k<t} \rho_k$, selectively dampens gradients from tokens following bad prefixes while preserving gradients from good prefixes. On AIME24 at staleness S=11, Prefix-Ratio GRPO achieves 0.500 avg@64, outperforming vanilla GRPO (0.400) by 10 percentage points. Selectivity analysis shows our method achieves 4.42× selectivity ratio, dampening 99.4% of bad-prefix tokens while only dampening 22.6% of good-prefix tokens.

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

1 INTRODUCTION

Reinforcement learning (RL) has emerged as a powerful paradigm for improving the reasoning capabilities of large language models (LLMs). Recent advances such as DeepSeek-R1 (DeepSeek-AI et al., 2025) and Qwen2.5 (Yang et al., 2024) have demonstrated remarkable performance on mathematical reasoning benchmarks through RL post-training with verifiable rewards. Scaling RL training to large models requires distributed systems that decouple rollout generation from learning (Xiao et al., 2026; Fu et al., 2025; Sheng et al., 2024), which introduces *staleness*: trajectories may be generated by a behavior policy that lags behind the current learner policy by multiple update steps.

Standard approaches to handling staleness use importance sampling with clipping (Schulman et al., 2015; 2017; Shao et al., 2024) to correct for distribution shift between the behavior and learner policies. However, these methods apply per-token importance ratios $\rho_t = \pi_\theta(y_t|x, y_{<t})/\pi_{\text{beh}}(y_t|x, y_{<t})$ that treat each token independently, ignoring the sequential dependencies inherent in autoregressive generation. In chain-of-thought reasoning, a flawed reasoning step early in a sequence (a "bad prefix") makes all subsequent tokens unreliable for learning, regardless of their individual token ratios. Standard per-token corrections fail to capture this—they may assign high gradient weight to tokens that follow flawed reasoning paths.

We propose Prefix-Ratio GRPO, which modifies the importance ratio to incorporate prefix information. Our key insight is that *if any prefix token has become unlikely under the current policy, all subsequent tokens should be downweighted*. We define a prefix-aware importance ratio $\tilde{\rho}_t = \underline{\rho}_t \cdot \rho_t$, where $\underline{\rho}_t = \min_{k<t} \rho_k$ is the minimum token ratio in the preceding prefix. This selectively dampens

[1] https://gitlab.com/fars-a/echo2-prefix-ratio-staleness

gradients from tokens following bad prefixes while preserving gradients from tokens following good prefixes.

Our contributions are:

- We propose Prefix-Ratio GRPO, a simple modification to GRPO that incorporates prefix information into importance ratios, enabling selective gradient dampening based on prefix quality.
- On AIME24 at staleness $S = 11$, Prefix-Ratio GRPO achieves 0.500 avg@64, outperforming vanilla GRPO (0.400, +10pp) and tight-clip GRPO (0.367, +13.3pp).
- We provide selectivity analysis showing Prefix-Ratio GRPO achieves 4.42× selectivity ratio, dampening 99.4% of bad-prefix tokens while only dampening 22.6% of good-prefix tokens.

2 Related Work

Reinforcement Learning for LLMs. Reinforcement learning from human feedback (RLHF) has emerged as a foundational paradigm for aligning large language models with human preferences (Kaufmann et al., 2023). Proximal Policy Optimization (PPO) (Schulman et al., 2017) has been widely adopted due to its stability through clipped surrogate objectives. To reduce computational overhead, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) eliminates the critic network by using group-relative advantages, enabling more efficient training. Direct Preference Optimization (DPO) (Rafailov et al., 2023) further simplifies alignment by directly optimizing on preference data without explicit reward modeling. Recent advances have demonstrated remarkable reasoning capabilities through RL post-training, with DeepSeek-R1 (DeepSeek-AI et al., 2025) and Qwen2.5 (Yang et al., 2024) achieving strong performance on mathematical reasoning benchmarks. DAPO (Yu et al., 2025) introduces dynamic advantage processing for improved stability, while REINFORCE++ (Hu et al., 2025) stabilizes critic-free optimization through global advantage normalization. VinePPO (Kazemnejad et al., 2024) refines credit assignment in RL training of LLMs through Monte Carlo tree search.

Off-Policy RL and Staleness Handling. Scaling RL training for LLMs requires distributed systems with asynchronous rollouts, which introduces staleness between rollout and training policies. Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) established the theoretical foundation for constraining policy updates, while PPO (Schulman et al., 2017) provides a practical approximation through clipping. Recent work has focused on managing staleness in distributed LLM training. ECHO-2 (Xiao et al., 2026) treats bounded policy staleness as a user-controlled parameter, enabling overlap between rollout generation and training. AReaL (Fu et al., 2025) and LlamaRL (Wu et al., 2025) develop large-scale asynchronous RL frameworks for efficient training. A-3PO (Li et al., 2025) approximates the proximal policy through interpolation to reduce computational overhead under staleness. StaleFlow (Li et al., 2026) jointly addresses data staleness and skewness through staleness-constrained rollout coordination. BAPO (Xi et al., 2025) identifies entropy collapse as a key failure mode in off-policy settings and proposes adaptive clipping to rebalance positive and negative gradient contributions. Zheng et al. (2025c) empirically characterize the "prosperity before collapse" phenomenon in off-policy RL for LLMs.

Prefix-Aware Methods. MinPRO (Lei et al., 2026) introduces prefix importance ratios for stabilizing policy optimization under off-policy conditions. The key insight is that the theoretically rigorous correction term is the prefix importance ratio rather than the token-level ratio, and relaxing it to a token-level approximation can induce instability. MinPRO replaces the cumulative prefix ratio with a non-cumulative surrogate based on the minimum token-level ratio in the preceding prefix. Our work applies similar intuition to gradient quality improvement: we use prefix-aware ratios to selectively dampen gradients from tokens following bad reasoning prefixes, improving downstream performance rather than focusing solely on training stability.

Mathematical Reasoning. Chain-of-thought prompting (Wei et al., 2022) demonstrated that LLMs can perform complex reasoning through step-by-step decomposition. Subsequent work has focused on improving mathematical reasoning through RL post-training (Wang et al., 2025).

DeepSeekMath (Shao et al., 2024) introduced GRPO for efficient RL training on mathematical reasoning tasks. Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025b) extends GRPO with sequence-level grouping for improved sample efficiency. Zheng et al. (2025a) provide a systematic study of stabilization techniques for RL with LLMs.

3 METHOD

3.1 PRELIMINARIES

We consider reinforcement learning with verifiable rewards (RLVR) for large language models, where each model output can be scored by an automatic verifier. Given a prompt x, the model generates a response $y = (y_1, y_2, \ldots, y_T)$ autoregressively, and receives a scalar reward r based on correctness.

Group Relative Policy Optimization. GRPO (Shao et al., 2024) eliminates the need for a separate value function by using group-relative advantages. For each prompt x, GRPO samples a group of G outputs $\{y^{(1)}, \ldots, y^{(G)}\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and computes advantages relative to the group mean reward. The GRPO objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\left(\rho_t\hat{A}_t, \text{clip}(\rho_t, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right) - \beta D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right] \quad (1)$$

where $\rho_t = \pi_\theta(y_t|x, y_{<t})/\pi_{\theta_{\text{old}}}(y_t|x, y_{<t})$ is the per-token importance ratio, $\hat{A}_t$ is the token-level advantage, ε is the clipping threshold, and β controls KL regularization.

Off-Policy Training and Staleness. In distributed RL systems, rollout generation is often decoupled from learning to improve throughput (Xiao et al., 2026; Fu et al., 2025). This introduces *staleness*: trajectories may be generated by a behavior policy π_{beh} that lags behind the current learner policy π_θ by up to S learner steps. The importance ratio then becomes $\rho_t = \pi_\theta(y_t|x, y_{<t})/\pi_{\text{beh}}(y_t|x, y_{<t})$, and clipping is applied to constrain updates when the policies diverge significantly.

3.2 THE PROBLEM: SEQUENTIAL DEPENDENCIES IN AUTOREGRESSIVE GENERATION

Standard GRPO uses per-token importance ratios ρ_t that treat each token independently. However, in autoregressive generation, the quality of token y_t depends critically on the quality of the preceding prefix $y_{<t}$. This creates a fundamental mismatch between the token-level correction and the sequential nature of language generation.

Prefix Importance Ratios. The theoretically correct importance sampling correction for off-policy policy gradients is the *prefix importance ratio* (Lei et al., 2026):

$$\rho_{1:t} = \frac{P_\theta(y_1, \ldots, y_t)}{P_{\text{beh}}(y_1, \ldots, y_t)} = \prod_{k=1}^{t}\rho_k \quad (2)$$

The token-level ratio ρ_t is an approximation that ignores the cumulative effect of prefix drift. When staleness is high and sequences are long, this approximation becomes increasingly unreliable.

Bad Prefixes Make Subsequent Tokens Unreliable. Consider a chain-of-thought reasoning trajectory where an early reasoning step is flawed. Under the current policy π_θ, this "bad prefix" has low probability—the model has learned to avoid such reasoning paths. However, the token-level ratio ρ_t for subsequent tokens may still be near 1 if those tokens are locally plausible given the (flawed) prefix. Standard GRPO would apply full gradient weight to these tokens, even though they follow unreliable reasoning. The key insight is that *if any prefix token has become unlikely under the current policy, all subsequent tokens should be downweighted*, as the entire reasoning path following a bad prefix is unreliable for learning.

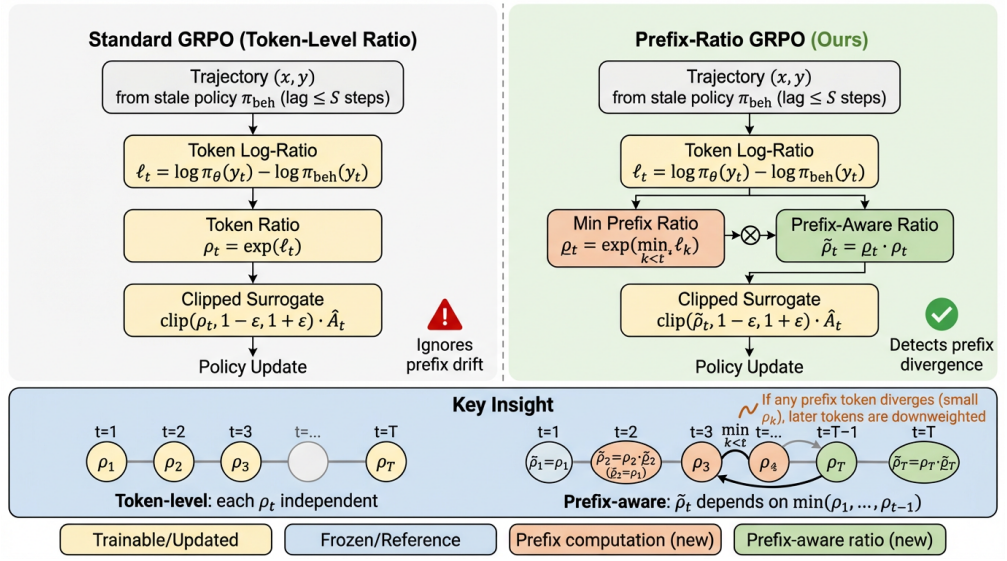

Figure 1: Comparison of Standard GRPO (left) and Prefix-Ratio GRPO (right). Standard GRPO applies per-token importance ratios ρ_t independently, while Prefix-Ratio GRPO uses prefix-aware ratios $\tilde{\rho}_t = \underline{\rho}_t \cdot \rho_t$ that selectively dampen tokens following bad prefixes.

3.3 Prefix-Ratio GRPO

We propose Prefix-Ratio GRPO, which modifies the importance ratio to incorporate prefix information while maintaining numerical stability.

Minimum Prefix Ratio. Following the intuition from MinPRO (Lei et al., 2026), we define the minimum prefix token ratio before position t:

$$\underline{\rho}_t = \min_{k<t} \rho_k = \exp\left(\min_{k<t} \ell_k\right) \tag{3}$$

where $\ell_k = \log \pi_\theta(y_k|x, y_{<k}) - \log \pi_{\text{beh}}(y_k|x, y_{<k})$ is the log-ratio at position k.

Prefix-Aware Importance Ratio. We replace the token-level ratio ρ_t with a prefix-aware surrogate:

$$\tilde{\rho}_t = \underline{\rho}_t \cdot \rho_t \tag{4}$$

This formulation ensures that if any earlier token has become unlikely under the current policy (small $\underline{\rho}_t$), later-token gradients are downweighted even if the local token ratio ρ_t is not extreme.

Prefix-Ratio GRPO Objective. The complete Prefix-Ratio GRPO objective replaces ρ_t with $\tilde{\rho}_t$ in the clipped surrogate:

$$\mathcal{J}_{\text{PR-GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\left(\tilde{\rho}_t\hat{A}_t, \text{clip}(\tilde{\rho}_t, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right) - \beta D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\right] \tag{5}$$

Figure 1 illustrates the difference between standard GRPO and Prefix-Ratio GRPO. In standard GRPO, each token's gradient weight depends only on its own ratio ρ_t. In Prefix-Ratio GRPO, the weight $\tilde{\rho}_t$ incorporates information about the entire prefix, enabling selective dampening of tokens following bad reasoning paths.

Table 1: Main results on AIME24 benchmark at staleness $S = 11$. Prefix-Ratio GRPO achieves the best performance (+10pp over vanilla, +13.3pp over tight-clip). All methods show pg_clipfrac=0.0 (clipping never activated). Best results in **bold**.

Method	ε	AIME24 avg@64	Problems Solved	pg_clipfrac	Seeds Collapsed
Vanilla GRPO	0.2	0.400	12/30	0.0	0/3
Tight-Clip GRPO	0.1	0.367	11/30	0.0	0/3
Prefix-Ratio GRPO	**0.2**	**0.500**	**15/30**	**0.0**	**0/3**

3.4 Implementation

Prefix-Ratio GRPO is a drop-in modification to standard GRPO that requires only access to behavior-policy token log-probabilities, which are commonly logged in RLHF/RLVR pipelines.

Computational Overhead. The additional computation is minimal: for each trajectory, we compute a cumulative minimum over log-ratios, which is $O(T)$ where T is the sequence length. This is negligible compared to the forward and backward passes through the model.

Integration with Bounded-Staleness Systems. Prefix-Ratio GRPO is compatible with bounded-staleness replay systems like ECHO-2 (Xiao et al., 2026). The modification is learner-side only: it does not require changes to rollout generation, reward computation, or the distributed rollout infrastructure. Trajectories are tagged with their behavior policy version, and the learner applies the prefix-aware ratio correction when computing gradients.

4 Experiments

4.1 Experimental Setup

Model and Training. We use Qwen3-8B (Yang et al., 2024) as the base model and train on the DAPO-Math-17K dataset for reinforcement learning with verifiable rewards. Training is conducted using the veRL framework (Sheng et al., 2024) in co-located mode with 8 A100-80GB GPUs. We simulate bounded staleness $S = 11$ by delaying policy snapshot publication every $\kappa = 10$ learner updates, following the ECHO-2 protocol (Xiao et al., 2026).

Methods. We compare three methods at staleness $S = 11$: (1) **Vanilla GRPO** with standard token-level clipping ($\varepsilon = 0.2$), (2) **Tight-Clip GRPO** with reduced clipping threshold ($\varepsilon = 0.1$) to test whether generic clipping strength suffices, and (3) **Prefix-Ratio GRPO** (ours) with prefix-aware importance ratios ($\varepsilon = 0.2$).

Evaluation. We evaluate on AIME24, a benchmark of 30 competition-level mathematics problems. Following standard practice, we report avg@64 accuracy (average accuracy over 64 samples per problem) and the number of problems solved (at least one correct answer out of 64). We run 3 seeds per method for stability testing (12 steps each), then extend the first surviving seed to 60 steps for quality evaluation.

4.2 Main Results

Table 1 presents the main experimental results at staleness $S = 11$. Prefix-Ratio GRPO achieves the best performance with AIME24 avg@64 of 0.500, outperforming vanilla GRPO (0.400, +10 percentage points) and tight-clip GRPO (0.367, +13.3 percentage points). Notably, Prefix-Ratio GRPO solves 15 out of 30 problems compared to 12 for vanilla and 11 for tight-clip, demonstrating qualitatively better reasoning capability.

All three methods remained stable throughout training with no seeds collapsing, contrary to prior reports of GRPO instability at $S = 11$ (Xiao et al., 2026). This may be attributed to the co-located training mode in veRL, which maintains tighter synchronization between rollout and training policies than fully distributed setups. Importantly, pg_clipfrac remained 0.0 throughout all experiments,

Table 2: Selectivity analysis of gradient dampening mechanisms. Prefix-Ratio GRPO achieves 2.9× higher selectivity than tight-clip, dampening 99.4% of bad-prefix tokens while only dampening 22.6% of good-prefix tokens. Best results in **bold**.

Method	Good-Prefix Dampen Rate	Bad-Prefix Dampen Rate	Selectivity Ratio
Vanilla GRPO	2.2%	5.1%	2.27
Tight-Clip GRPO	9.9%	15.1%	1.52
Prefix-Ratio GRPO	**22.6%**	**99.4%**	**4.42**

indicating that the clipping mechanism was never activated at this operating point. This suggests that the performance improvement from Prefix-Ratio GRPO operates through gradient quality enhancement rather than through its interaction with clipping.

4.3 Selectivity Analysis

To understand the mechanism behind Prefix-Ratio GRPO's improvement, we analyze the selectivity of gradient dampening across methods. We define a token as "dampened" if its effective importance ratio falls below a threshold, and measure the dampen rate separately for tokens following good prefixes versus bad prefixes.

Table 2 shows the selectivity analysis results. Prefix-Ratio GRPO achieves a selectivity ratio of 4.42, meaning it dampens bad-prefix tokens 4.42× more frequently than good-prefix tokens. In contrast, tight-clip GRPO achieves only 1.52 selectivity ratio, and vanilla GRPO achieves 2.27. Critically, Prefix-Ratio GRPO dampens 99.4% of bad-prefix tokens while only dampening 22.6% of good-prefix tokens, demonstrating highly selective gradient filtering.

The selectivity analysis reveals why Prefix-Ratio GRPO outperforms tight-clip GRPO despite both methods dampening gradients. Tight-clip GRPO applies uniform dampening regardless of prefix quality, suppressing both useful and unreliable gradients, whereas Prefix-Ratio GRPO selectively targets tokens following bad reasoning paths.

4.4 Training Dynamics

Figure 2 shows the training reward curves for all three methods. Despite achieving significantly different downstream performance, all methods exhibit similar reward improvement trajectories during training. The 5-step moving average reward improves from approximately -0.998 at step 5 to around -0.55 to -0.68 by step 55-56 across all methods.

The similar training dynamics but different downstream performance supports our hypothesis that Prefix-Ratio GRPO improves gradient quality rather than training stability. The prefix-aware importance ratio selectively filters unreliable gradients from tokens following bad reasoning prefixes, leading to better policy updates even when aggregate training metrics appear similar.

4.5 Discussion

Stability Hypothesis. Our experiments did not observe the training collapse at $S = 11$ reported by ECHO-2 (Xiao et al., 2026). All three methods remained stable with KL divergence in a narrow band ($\sim$0.01) and no numerical failures. This may be due to differences between our co-located veRL setup and ECHO-2's fully distributed architecture. Consequently, the stability benefit of Prefix-Ratio GRPO remains inconclusive in our experiments.

Gradient Quality Mechanism. The performance improvement from Prefix-Ratio GRPO appears to operate through gradient quality enhancement rather than stability. The key evidence is: (1) clipping was never activated (pg_clipfrac=0.0); (2) training dynamics are nearly identical across methods; (3) the selectivity analysis demonstrates highly selective gradient filtering. This selective dampening improves the quality of policy updates, leading to better downstream performance.

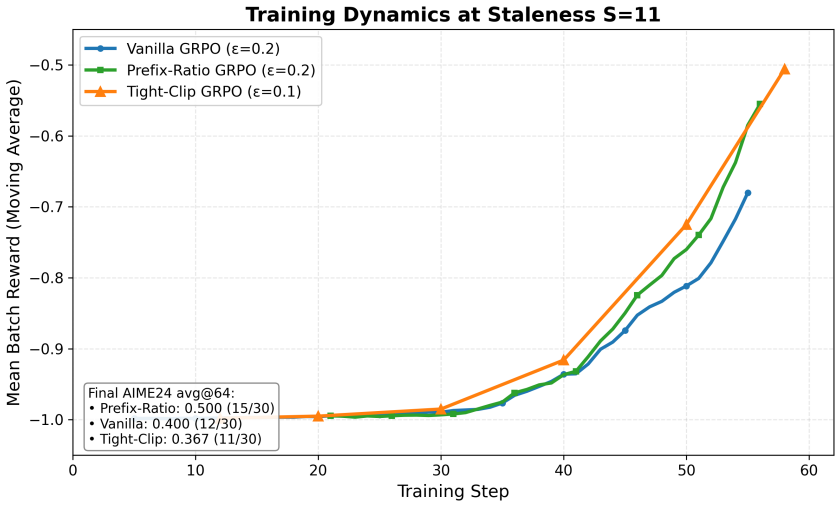

Figure 2: Training dynamics at staleness $S = 11$. All three methods show similar reward improvement trajectories, but Prefix-Ratio GRPO achieves the best downstream AIME24 performance (0.500) despite similar training rewards.

5 Conclusion

We presented Prefix-Ratio GRPO, a simple modification to GRPO that incorporates prefix information into importance ratios for off-policy reinforcement learning with LLMs. By using prefix-aware ratios $\tilde{\rho}_t = \underline{\rho}_t \cdot \rho_t$, our method selectively dampens gradients from tokens following bad reasoning prefixes while preserving gradients from tokens following good prefixes. On AIME24 at staleness $S = 11$, Prefix-Ratio GRPO achieves 0.500 avg@64, outperforming vanilla GRPO by 10 percentage points and tight-clip GRPO by 13.3 percentage points, with 4.42× selectivity in gradient dampening. The stability hypothesis remains inconclusive as vanilla GRPO did not collapse in our setup. Future work includes evaluating at higher staleness regimes where collapse may occur, extending to other reasoning domains, and theoretical analysis of prefix-ratio dynamics.

References

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, Y. 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, Xu Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo 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.

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

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

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. 2024.

Shiye Lei, Zhihao Cheng, and Dacheng Tao. A step back: Prefix importance ratio stabilizes policy optimization. 2026.

Haoyang Li, Sheng Lin, Fangcheng Fu, Yuming Zhou, Xiaodong Ji, Yanfeng Zhao, Lefeng Wang, Jie Jiang, and Bin Cui. Unleashing efficient asynchronous rl post-training via staleness-constrained rollout coordination. *ArXiv*, abs/2601.12784, 2026.

Xiaocan Li, Shiliang Wu, and Zheng Shen. A-3po: Accelerating asynchronous llm training with staleness-aware proximal policy approximation. *ArXiv*, abs/2512.06547, 2025.

Rafael Rafailov, Archit Sharma, E. Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290, 2023.

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.

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.

Pengyuan Wang, Tian-Shuo Liu, Chenyang Wang, Yidi Wang, Shuo Yan, Cheng-Xing Jia, Xu-Hui Liu, Xin-Wei Chen, Jiachen Xu, Ziniu Li, and Yang Yu. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, Xia Tang, Yundi Qian, Beibei Zhu, and R. Hou. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm training. *ArXiv*, abs/2505.24034, 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.

Jie Xiao, Meng Chen, Qingnan Ren, Jingwei Song, Jiaqi Huang, Yangshen Deng, Chris Tong, Wanyi Chen, Suli Wang, Ziqian Bi, Shuo Lu, Yiqun Duan, Xu Wang, Rymon Yu, Ween Yang, Lynn Ai, Eric Yang, and Bill Shi. Echo-2: A large-scale distributed rollout framework for cost-efficient reinforcement learning. 2026.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

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, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Jun Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, An Yang, Jingren Zhou, and Junyang Lin. Stabilizing reinforcement learning with llms: Formulation and practices. *ArXiv*, abs/2512.01374, 2025a.

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. *ArXiv*, abs/2507.18071, 2025b.

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