TRAJECTORY-AWARE POLICY OPTIMIZATION: CLOSING THE STATE DISTRIBUTION GAP IN LONG-HORIZON LLM RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has become a dominant paradigm for enhancing LLMs' reasoning capabilities. Yet long-horizon autoregressive rollouts introduce a fundamental limitation: local policy mismatch can compound into severe state distribution shift over trajectories. Existing LLM RL methods, such as GRPO and GSPO, control policy mismatch at the token or sequence level, but overlook this trajectory-induced shift, causing training instability and eventual collapse as the horizon grows. To close the state distribution gap, we propose *Trajectory-Aware Policy Optimization (TAPO)*, a principled framework that theoretically derives the state-distribution correction from the off-policy policy gradient and incorporates it into the RL objective. TAPO further employs a geometric-mean estimator for variance reduction, together with a length-aware dynamic clipping that prevents overly deviated updates across the trajectory. Extensive experiments demonstrate that TAPO significantly surpasses existing LLM RL algorithms across eight competition-level benchmarks, achieving up to 48% and 12% improvement over GRPO on mathematical reasoning and algorithmic coding, with substantially enhanced training stability.

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a critical paradigm for eliciting complex reasoning capabilities in Large Language Models (LLMs) (Guo et al., 2025; Comanici et al., 2025; OpenAI, 2024), yielding major breakthroughs across long-horizon reasoning frontiers including mathematical problem-solving (Shao et al., 2024; Wang et al., 2024; Luo et al., 2025), algorithmic coding (Dou et al., 2024; Anthropic, 2024), and deep search (OpenAI, 2025; Jin et al., 2025).

However, long-horizon autoregressive rollouts expose RL to a fundamental vulnerability: local policy mismatch can compound into severe state distribution shift over the trajectory. At each step, the target policy acts on a state induced by all previously sampled tokens, making its state distribution dependent on the entire history of policy deviations. Thus, even modest deviations at individual positions can propagate through the trajectory, progressively driving the target policy toward states increasingly distant from the behavior policy's state distribution. As empirically demonstrated in Fig. 1, while the local token-level ratio remains relatively controllable across positions, the state-level ratio progressively drifts away from unity over long horizons.

Importance sampling provides a principled mechanism for correcting off-policy sampling mismatch, yet existing LLM RL methods primarily perform this correction at the action level. Two representative approaches, GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025), operate at different granularities: GRPO applies token-level importance ratios, while GSPO regards the full sequence as an action and aggregates token-level ratios into a sequence-level correction. However, neither explicitly corrects for the state distribution shift induced by the preceding trajectory. Consequently, as response lengths grow during training, the accumulated state shift manifests as an increasing KL divergence between the behavior and target policies, eventually driving training into collapse, as shown in Fig. 2. This exposes a fundamental gap in existing policy optimization: controlling action-level mismatch at the token or sequence level does not guarantee control of the state distribution shift induced over long-horizon trajectories.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

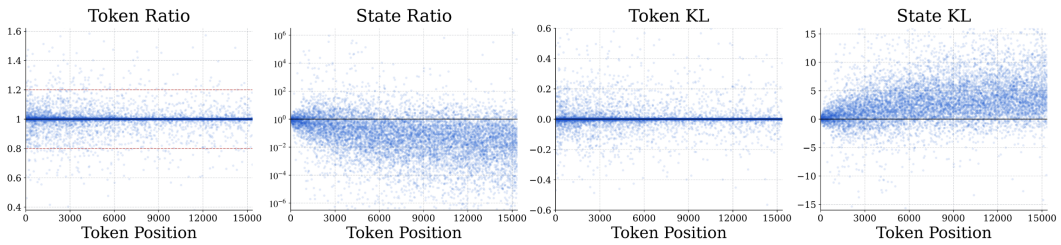


Figure 1: Policy shifts along long-horizon token positions, measured by token- and state-level policy ratios and KL divergences. The dots are randomly sampled through the training process of GRPO. While the token-level ratio remains relatively stable across positions, the state-level ratio progressively deviates from unity toward later positions, revealing the accumulation of policy shift over long-horizon trajectories.

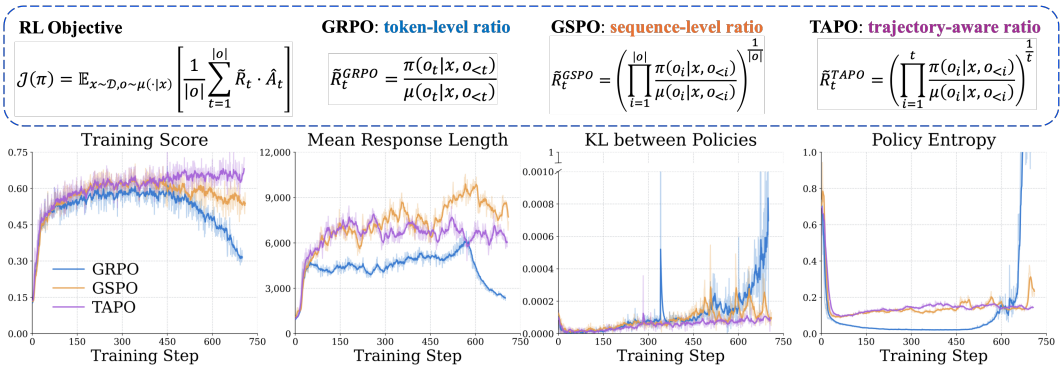


Figure 2: Formulations (upper) and training dynamics (bottom) of GRPO, GSPO, and TAPO. GRPO and GSPO apply policy-ratio correction at the token and sequence levels, respectively, whereas TAPO explicitly incorporates the state-distribution correction induced by the preceding trajectory. As response lengths increase, GRPO and GSPO exhibit growing KL divergence and eventual training collapse, while TAPO maintains controlled KL divergence and stable improvement.

To close this gap, we propose *Trajectory-Aware Policy Optimization* (TAPO), a theoretically grounded RL algorithm that fundamentally corrects the state distribution shift over the trajectory. We first derive the off-policy policy gradient to reveal the state-distribution correction required for long-horizon autoregressive trajectories, and incorporate this correction into the RL objective. To control the variance of the state-level correction, we use a geometric-mean estimator over the preceding trajectory, yielding a stable and length-robust trajectory-aware policy ratio (Fig 2). Building on this formulation, we further introduce a dynamic clipping mechanism that adapts the clipping boundary to the length of preceding trajectory, preventing overly aggressive updates as the horizon grows while preserving sufficient flexibility at early explorations.

We conduct extensive experiments across various LLMs and competition-level benchmarks for mathematical reasoning and algorithmic coding. TAPO consistently outperforms existing RL algorithms, achieving up to 48% and 12% average improvements over GRPO on math and code, respectively, demonstrating substantial gains in long-horizon reasoning performance. Moreover, as Fig. 2 shows, while GRPO and GSPO eventually suffer from training collapse as the policy shift grows, TAPO maintains a steadily increasing training score with controlled KL divergence. These results demonstrate that explicitly correcting the trajectory-induced state distribution shift not only improves reasoning performance but also substantially enhances training stability.

In summary, our contributions are threefold:

- We identify and formalize the accumulation of state distribution shift as a fundamental issue in long-horizon LLM RL, and derive its corresponding correction from the off-policy policy gradient.
- We propose TAPO, a theoretically grounded RL algorithm, employing geometric-mean trajectory-aware ratio together with length-aware dynamic clipping for stable long-horizon optimization.

- Extensive experiments across mathematical reasoning and algorithmic coding demonstrate that TAPO consistently improves reasoning performance with enhanced training stability.

2 PRELIMINARY

2.1 MDP FORMULATION FOR AUTOREGRESSIVE LLMS

The autoregressive generation of an LLM can be formulated as a finite-horizon Markov decision process (MDP). Given a prompt x, at step t, the state $s_t = (x, o_{<t})$ is defined by prompt together with the preceding generated tokens $o_{<t} = (o_1, \ldots, o_{t-1})$, and the action $o_t \sim \pi(\cdot \mid s_t)$ is the t-th token sampled from the policy π (parameterized by θ). A trajectory is therefore given by $\tau = (s_1, o_1, \ldots, o_T, s_{T+1}) \sim P^\pi(\cdot|x)$, with the next state deterministically induced by the current action as $s_{t+1} = (x, o_{\le t})$. Thus, the terminal state is $s_{T+1} = (x, o)$, where $o = (o_1, \ldots, o_T)$ denotes the complete response.

Typically, LLM reasoning tasks provide only an outcome reward $R(\tau)$ for the entire response trajectory. The objective of RL is to maximize the expected reward over generated trajectories:

$$\mathcal{J}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim P^\pi(\cdot|x)}\left[R(\tau)\right], \qquad \nabla_\theta \mathcal{J}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim P^\pi(\cdot|x)}\left[\sum_{t=1}^{T} \nabla_\theta \log \pi(o_t \mid s_t) A_t\right] \tag{1}$$

where the gradient is obtained by the policy gradient theorem (Sutton et al., 1999), and A_t denotes the advantage assigned to the token o_t.

In practice, LLM RL methods typically sample the trajectories from the behavior policy μ while optimize the target policy π, yielding an off-policy mismatch that requires correction.

2.2 GRPO: TOKEN-LEVEL CORRECTION

Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025) follows PPO (Schulman et al., 2017) in applying an action-level importance ratio r_t^g to individual tokens to correct the off-policy mismatch. Specifically, the GRPO objective is given by:

$$\mathcal{J}_{\text{GRPO}}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\{o^g\}_{g=1}^G\sim\mu(\cdot|x)}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o^g|}\sum_{t=1}^{|o^g|}\min\left(r_t^g\hat{A}_t^g, \text{clip}\left(r_t^g, 1-\epsilon^l, 1+\epsilon^h\right)\hat{A}_t^g\right)\right] \tag{2}$$
$$\text{where} \quad r_t^g = \frac{\pi(o_t^g \mid x, o_{<t}^g)}{\mu(o_t^g \mid x, o_{<t}^g)}$$

GRPO estimates advantages in a group-relative manner, where the behavior policy μ samples a group of G individual responses $\{o^g\}_{g=1}^G$ for a given query, and the advantage assigned to the g-th response is obtained by normalizing its reward R^g within the group, thus $\hat{A}_t^g = \frac{R^g - \text{mean}(\{R^g\}_{g=1}^G)}{\text{std}(\{R^g\}_{g=1}^G)}$. Moreover, subsequent GRPO variants, such as DAPO (Yu et al., 2025), DCPO (Yang et al., 2025b), DPPO (Qi et al., 2026), and RIPO (Cai et al., 2026) maintain the token-level importance ratio.

2.3 GSPO: SEQUENCE-LEVEL CORRECTION

Group Sequence Policy Optimization (GSPO) (Zheng et al., 2025) instead regards the entire response as a single action to match the unit of sequence-level reward, and applies importance-ratio correction at the sequence level. Specifically, GSPO employs the following optimization objective:

$$\begin{aligned}\mathcal{J}_{\text{GSPO}}(\pi) &= \mathbb{E}_{x\sim\mathcal{D},\{o^g\}_{g=1}^G\sim\mu(\cdot|x)}\left[\frac{1}{G}\sum_{g=1}^{G}\min\left(s^g\hat{A}^g, \text{clip}\left(s^g, 1-\epsilon^l, 1+\epsilon^h\right)\hat{A}^g\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D},\{o^g\}_{g=1}^G\sim\mu(\cdot|x)}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o^g|}\sum_{t=1}^{|o^g|}\min\left(s_t^g\hat{A}_t^g, \text{clip}\left(s_t^g, 1-\epsilon^l, 1+\epsilon^h\right)\hat{A}_t^g\right)\right] \end{aligned} \tag{3}$$
$$\text{where} \quad s_t^g = \text{sg}[s^g]\frac{\pi(o_t^g \mid x, o_{<t}^g)}{\text{sg}[\pi(o_t^g \mid x, o_{<t}^g)]}, \quad s^g = \left(\prod_{i=1}^{|o^g|}\frac{\pi(o_i^g \mid x, o_{<i}^g)}{\mu(o_i^g \mid x, o_{<i}^g)}\right)^{\frac{1}{|o^g|}}$$

$sg[\cdot]$ denotes only taking the numerical value but stopping the gradient, s^g is the sequence-level likelihood of importance ratio, and the term $\frac{\pi(o_t^g|x,o_{<t}^g)}{\text{sg}[\pi(o_t^g|x,o_{<t}^g)]}$ with identity value is used to assign the gradient to the token at position t.

However, GSPO does not fully align with the autoregressive MDP structure of LLM RL, where each token is generated as an action at a state induced by the preceding trajectory. Treating the entire trajectory as the unit of clipping also leads to low token efficiency: once the sequence-level ratio falls outside the clipping range, all tokens in the trajectory are clipped, including intermediate tokens whose policy mismatch may remain well controlled. Conversely, even when the sequence-level ratio remains within the clipping range, substantial state distribution shifts may still occur at intermediate positions. Although these local deviations can be averaged out in the final sequence-level ratio, the corresponding token-level updates are still influenced by these severely shifted states, potentially causing unstable optimization.

Moreover, despite their different correction granularities, both GRPO and GSPO correct policy mismatch through action likelihoods, without explicitly accounting for the state distribution induced by the preceding trajectory. We next show that a principled off-policy policy gradient necessarily involves an additional state-distribution correction induced by the preceding trajectory.

3 METHOD

In this section, we propose Trajectory-Aware Policy Optimization (TAPO) to correct the state distribution shift induced by preceding trajectory. We first derive the off-policy policy gradient to establish the theoretical importance ratio, then develop a variance-reduced correction and a length-aware dynamic clipping for stable training, and finally integrate these components into the unified TAPO objective.

3.1 FROM ACTION CORRECTION TO STATE CORRECTION

For a target policy π optimized using trajectories sampled from a behavior policy μ, the objective can be equivalently evaluated under the behavior policy via trajectory-level importance sampling:

$$\mathcal{J}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim P^{\pi}(\cdot|x)}\left[R(\tau)\right] = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim P^{\mu}(\cdot|x)}\left[\left(\frac{P^{\pi}(\tau|x)}{P^{\mu}(\tau|x)}\right)R(\tau)\right] \tag{4}$$

The resulting off-policy policy gradient is given by the following theorem. [1]

Theorem 1 (Off-Policy Policy Gradient)**.** *Given target policy π to be optimized and behavior policy μ for sampling trajectories, the off-policy policy gradient is*

$$\nabla_\theta \mathcal{J}(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{t=1}^{T}\mathbb{E}_{s_t\sim d_t^{\mu}(\cdot|x),\,o_t\sim\mu(\cdot|s_t)}\left[\frac{d_t^{\pi}(s_t|x)}{d_t^{\mu}(s_t|x)}\frac{\pi(o_t|s_t)}{\mu(o_t|s_t)}\nabla_\theta\log\pi(o_t|s_t)A^{\pi}(s_t,o_t)\right]\right] \tag{5}$$

where $A^{\pi}(s_t,o_t)$ is the advantage function, $d_t^{\pi}(s_t|x) \triangleq \Pr_{\pi}(s_t|x) = \sum_{\tau_{<t}:s_t} P^{\pi}(\tau_{<t}|x)$ and $d_t^{\mu}(s_t|x) \triangleq \Pr_{\mu}(s_t|x) = \sum_{\tau_{<t}:s_t} P^{\mu}(\tau_{<t}|x)$ denote the state distributions induced by π and μ at step t, respectively.

The theorem reveals that the exact off-policy correction consists of two components: a state-distribution correction $\frac{d_t^{\pi}(s_t|x)}{d_t^{\mu}(s_t|x)}$ and an action-level correction $\frac{\pi(o_t|s_t)}{\mu(o_t|s_t)}$. In traditional RL, policy optimization is typically formulated in continuing or infinite-horizon MDPs, where the state distributions are characterized by a single MDP-induced stationary distribution (Schulman et al., 2016; 2017). Under this formulation, state-distribution mismatch is handled implicitly through the policy-induced stationary distribution rather than explicitly corrected at each time step, and practical off-policy methods therefore often focus on action-level importance ratios.

In contrast, autoregressive LLM generation forms a finite-horizon, tree-structured MDP, where each generated token extends the state and thereby determines the branch of all subsequent states. There

[1] The detailed derivation is deferred to Appendix.

is no single stationary state distribution that characterizes the states encountered at different token positions. Consequently, local policy mismatch at preceding positions accumulates along the trajectory and can induce substantial state-distribution mismatch at later positions, making the state-distribution correction non-negligible for long-horizon autoregressive generation.

Specifically, the state transition of LLM RL is deterministic, each state $s_t = (x, o_{<t})$ uniquely identifies the preceding token sequence, thus the state-distribution ratio of LLM reduces to the likelihood ratio of the preceding trajectory:

$$\frac{d_t^\pi(s_t|x)}{d_t^\mu(s_t|x)}A = \frac{\sum_{\tau_{<t}:s_t} P^\pi(\tau_{<t}|x)}{\sum_{\tau_{<t}:s_t} P^\mu(\tau_{<t}|x)} = \frac{\prod_{i=1}^{t-1}\pi(o_i|s_i)}{\prod_{i=1}^{t-1}\mu(o_i|s_i)} = \prod_{i=1}^{t-1}\frac{\pi(o_i|s_i)}{\mu(o_i|s_i)}. \tag{6}$$

Consequently, the theoretically required importance correction $\mathcal{R}_t$ at token t can be written as

$$\mathcal{R}_t \triangleq \underbrace{\frac{d_t^\pi(s_t|x)}{d_t^\mu(s_t|x)}}_{\text{state correction}} \cdot \underbrace{\frac{\pi(o_t \mid s_t)}{\mu(o_t \mid s_t)}}_{\text{action correction}} = \prod_{i=1}^{t}\frac{\pi(o_i \mid s_i)}{\mu(o_i \mid s_i)}. \tag{7}$$

3.2 Trajectory-Aware Policy Optimization

Based on the theoretically required importance correction derived above, we first formulate a trajectory-aware corrected surrogate objective for LLM RL:

$$\mathcal{J}_{traj}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\{o^g\}_{g=1}^G\sim\mu(\cdot|x)}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o^g|}\sum_{t=1}^{|o^g|}\min\left(\mathcal{R}_t^g\hat{A}_t^g, \text{clip}\left(\mathcal{R}_t^g, 1-\epsilon^l, 1+\epsilon^h\right)\hat{A}_t^g\right)\right] \tag{8}$$

While the ratio $\mathcal{R}_t^g = \prod_{i=1}^{t}\frac{\pi(o_i^g|s_i^g)}{\mu(o_i^g|s_i^g)}$ provides the theoretically required trajectory-aware correction, directly using the product of per-token likelihood ratios can lead to high variance and unstable optimization, especially for long-horizon generation. Moreover, both the variance and the effective range of $\mathcal{R}_t^g$ vary with the length of the preceding trajectory, making the fixed clipping boundaries used by GRPO and GSPO inadequate. We therefore introduce two complementary mechanisms to make trajectory-aware correction practical: variance-controlled trajectory correction and length-aware dynamic clipping.

3.2.1 Variance-Controlled Trajectory Correction

The correction ratio $\mathcal{R}_t$ is a product of t per-token likelihood ratios, whose variance can grow rapidly with trajectory length, thus making the training unstable. To control this variance while preserving the trajectory-level correction structure, we replace the product with its geometric mean (we omit the superscript g for clarity):

$$\widetilde{\mathcal{R}}_t \triangleq \left(\prod_{i=1}^{t}\frac{\pi(o_i \mid s_i)}{\mu(o_i \mid s_i)}\right)^{1/t} = \mathcal{R}_t^{1/t}. \tag{9}$$

The variance reduction induced by geometric averaging can be characterized as follows.

Theorem 2 (Variance of Trajectory Correction Ratio)**.** *Let $r_i = \frac{\pi(o_i|s_i)}{\mu(o_i|s_i)}$ and suppose $\log r_i$ are i.i.d. Gaussian with variance σ^2. Since $\mathbb{E}_\mu[r_i] = 1$, we have $\log r_i \sim \mathcal{N}(-\frac{\sigma^2}{2}, \sigma^2)$. Then the exact trajectory-level correction and its geometric-mean counterpart satisfy*

$$\text{Var}_\mu(\mathcal{R}_t) = e^{t\sigma^2} - 1, \qquad \text{Var}_\mu(\widetilde{\mathcal{R}}_t) = e^{-\sigma^2+\sigma^2/t}\left(e^{\sigma^2/t} - 1\right). \tag{10}$$

Hence, the variance of the product correction grows exponentially with the length of preceding trajectory, whereas the variance of the geometric-mean correction decays as $O(1/t)$.

3.2.2 Length-Aware Dynamic Clipping

Geometric averaging resolves the variance explosion of the exact trajectory correction, but it also makes the correction increasingly concentrated as the trajectory grows. In particular, the distribution

of $\widetilde{\mathcal{R}}_t$ becomes progressively narrower in log-space, implying that a fixed clipping boundary cannot provide a consistent level of tolerated proximal region across different token positions. We therefore derive a length-aware clipping rule from this concentration property.

Theorem 3 (Concentration of Geometric-Mean Correction)**.** *Let $X_i = \log r_i$ be independent random variables such that $X_i - \mathbb{E}[X_i]$ is σ^2-sub-Gaussian, i.e.,*

$$\mathbb{E}\left[\exp\left(\lambda(X_i - \mathbb{E}[X_i])\right)\right] \le \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \qquad \forall\lambda \in \mathbb{R}. \tag{11}$$

For the geometric-mean correction $\widetilde{\mathcal{R}}_t = \exp\left(\frac{1}{t}\sum_{i=1}^{t} X_i\right)$, its log-ratio satisfies

$$\Pr\left(\left|\log\widetilde{\mathcal{R}}_t - \mathbb{E}[\log\widetilde{\mathcal{R}}_t]\right| \ge \epsilon\right) \le 2\exp\left(-\frac{t\epsilon^2}{2\sigma^2}\right). \tag{12}$$

Thus, for a desired coverage probability $1-\delta$, the deviation radius in log-space scales as $\epsilon_t = \sigma\sqrt{\frac{2\log(2/\delta)}{t}} = O\left(\frac{1}{\sqrt{t}}\right)$. Therefore, the effective range of $\widetilde{\mathcal{R}}_t$ contracts with trajectory length: early tokens require a wider tolerance, whereas later tokens admit a tighter clipping range. This perspective is consistent with the clipping boundaries adopted by existing methods: GRPO uses clipping boundaries on the order of 10^{-1} for token-level ratios, whereas GSPO uses substantially tighter boundaries on the order of 10^{-4} for sequence-level ratios.

Theorem 3 motivates us to adopt a length-aware clipping boundary whose length-dependent scaling follows $1/\sqrt{t}$. Specifically,

$$\epsilon_t^l = \epsilon_{\min}^l + (\epsilon_{\max}^l - \epsilon_{\min}^l)\frac{1/\sqrt{t} - 1/\sqrt{T}}{1 - 1/\sqrt{T}}, \qquad \epsilon_t^h = \epsilon_{\min}^h + (\epsilon_{\max}^h - \epsilon_{\min}^h)\frac{1/\sqrt{t} - 1/\sqrt{T}}{1 - 1/\sqrt{T}}. \tag{13}$$

The endpoint normalization maps the $1/\sqrt{t}$ scaling to the interval $[1, T]$, ensuring that the clipping range starts from a relatively permissive boundary $\epsilon_{\max}$ and gradually tightens toward a conservative boundary $\epsilon_{\min}$ as the trajectory grows.

3.2.3 Trajectory-Aware Policy Optimization

Combining variance-controlled trajectory correction with length-aware dynamic clipping, we obtain the final TAPO objective:

$$\mathcal{J}_{\text{TAPO}}(\pi) = \mathbb{E}_{x\sim\mathcal{D},\{o^g\}_{g=1}^{G}\sim\mu(\cdot|x)}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|o^g|}\sum_{t=1}^{|o^g|}\min\left(\widehat{\mathcal{R}}_t^g\hat{A}_t^g, \text{clip}\left(\widehat{\mathcal{R}}_t^g, 1-\epsilon_t^l, 1+\epsilon_t^h\right)\hat{A}_t^g\right)\right]$$
$$\text{where}\quad \widehat{\mathcal{R}}_t^g = \text{sg}\left[\widetilde{\mathcal{R}}_t^g\right]\frac{\pi(o_t^g \mid x, o_{<t}^g)}{\text{sg}\left[\pi(o_t^g \mid x, o_{<t}^g)\right]}, \quad \widetilde{\mathcal{R}}_t^g = \left(\prod_{i=1}^{t}\frac{\pi(o_i^g \mid x, o_{<i}^g)}{\mu(o_i^g \mid x, o_{<i}^g)}\right)^{1/t}. \tag{14}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, ϵ_t^l and ϵ_t^h follow the length-aware schedule defined in Eq. 13. Although $\widehat{\mathcal{R}}_t^g$ has the same forward value as $\widetilde{\mathcal{R}}_t^g$, its gradient satisfies

$$\nabla_\theta\widehat{\mathcal{R}}_t^g = \widetilde{\mathcal{R}}_t^g\nabla_\theta\log\pi(o_t^g \mid x, o_{<t}^g), \tag{15}$$

which decouples the trajectory-level correction from the gradient of preceding actions and retains the local policy-gradient structure of the current action. This construction is consistent with the off-policy policy gradient derived in Theorem 1: the trajectory-level ratio corrects the accumulated distribution shift, while the current action likelihood provides the policy-gradient direction.

By comparison, GRPO is action-corrected but state-agnostic, using only the local likelihood ratio for each token, whereas GSPO applies a single sequence-level correction uniformly across all tokens. TAPO instead provides trajectory-aware, token-specific correction, weighting each token according to the accumulated policy shift along the trajectory leading to its current state. Consequently, TAPO offers a principled middle ground between local action correction and global sequence correction, explicitly bridging the state-distribution gap in long-horizon autoregressive generation and enabling consistently stable training, as demonstrated in Fig. 2.

Table 1: Comparison of different RL algorithms trained on Qwen3-8B-Base across various mathematical reasoning and algorithmic coding benchmarks (evaluated by Avg@8). We highlight each cell as best, second best, and third best.

Mathematical Reasoning						
Method	Granularity	AIME24	AIME25	BRUMO25	SMT25	Average
Base	none	3.8	3.3	15.0	10.8	8.2 (-70.4%)
GRPO	token-level	34.6	20.8	33.3	21.9	27.7 (+0.0%)
DAPO	token-level	45.0	27.5	37.5	28.5	34.6 (+24.9%)
DCPO	token-level	36.3	27.5	42.8	31.6	34.6 (+24.9%)
SAPO	token-level	37.3	28.7	41.7	32.3	35.0 (+26.4%)
CISPO	token-level	31.7	23.8	38.3	28.0	30.5 (+10.1%)
Clip-Cov	token-level	36.3	22.9	36.6	30.2	31.5 (+13.7%)
DPPO	token-level	46.7	32.5	42.5	34.2	39.0 (+40.8%)
RIPO	token-level	47.9	32.5	47.5	34.2	40.5 (+46.2%)
GSPO	sequence-level	42.9	28.7	36.3	32.5	35.1 (+26.7%)
GMPO	sequence-level	41.7	30.0	35.8	33.0	35.1 (+26.7%)
TAPO	trajectory-aware	49.6	33.3	47.2	33.8	41.0 (+48.0%)
Code Generation						
Method	Granularity	Codeforces	CodeContests	TACO	APPS	Average
Base	none	32.3	32.1	16.5	32.1	28.3 (-45.9%)
GRPO	token-level	61.9	57.4	36.2	54.2	52.4 (+0.0%)
DAPO	token-level	64.0	59.1	37.4	56.8	54.3 (+3.6%)
DCPO	token-level	63.5	58.7	36.8	57.2	54.1 (+3.2%)
SAPO	token-level	63.7	57.0	37.5	55.9	53.5 (+2.1%)
CISPO	token-level	61.6	54.6	35.1	53.9	51.3 (-2.1%)
Clip-Cov	token-level	62.9	57.4	36.0	55.5	53.0 (+1.2%)
DPPO	token-level	63.5	58.6	37.6	55.6	53.8 (+2.7%)
RIPO	token-level	68.2	62.3	39.3	58.9	57.2 (+9.2%)
GSPO	sequence-level	63.8	57.9	36.6	56.3	53.7 (+2.5%)
GMPO	sequence-level	64.0	59.2	37.4	56.8	54.4 (+3.8%)
TAPO	trajectory-aware	70.2	64.5	40.7	58.9	58.6 (+11.8%)

4 EXPERIMENTS

To validate the effectiveness of TAPO, we conduct extensive experiments across eight competition-level benchmarks spanning both mathematical reasoning and algorithmic coding, with comparisons against ten representative RL algorithms on models of different scales.

4.1 EXPERIMENTAL SETUP

Baselines. We compare the proposed TAPO to ten representative RL algorithms spanning different correction granularities of policy shift. For token-level correction, we compare against GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), DCPO (Yang et al., 2025b), SAPO (Gao et al., 2025), CISPO (Chen et al., 2025), Clip-Cov (Cui et al., 2025b), DPPO (Qi et al., 2026), and RIPO (Cai et al., 2026) ; for sequence-level correction, we compare against GSPO (Zheng et al., 2025) and GMPO (Zhao et al., 2026). All compared methods follow their default hyperparameter settings.

Following common practice (Ye et al., 2020; Cai et al., 2026), we apply dual clipping to the importance ratio with lower and upper bounds of 0.5 and 10, and remove the KL penalty term for all methods (Yu et al., 2025; Cai et al., 2026). All algorithms adopt the GRPO-style group-relative advantage estimator, and trained on Qwen3-8B-Base (Yang et al., 2025a). For TAPO, we set $\epsilon^l_{\max} = \epsilon^h_{\max} = 0.2$ and $\epsilon^l_{\min} = 0.0003$, $\epsilon^h_{\min} = 0.0004$, respectively.

Training. We evaluate the improvement induced by each RL algorithm on two complementary domains, mathematical reasoning and algorithmic coding, both of which admit verifiable rewards and are widely regarded as strong indicators of general reasoning ability. For mathematical reasoning we train on DAPO-Math-17k (Yu et al., 2025), and for algorithmic coding on Eurus-Code-25k (Cui et al., 2025a). For each query we generate 8 rollouts (namely, the group size is 8) with a maximum response length of 20k tokens for thorough reasoning. In each RL iteration, the behavior policy μ totally generates 1024 rollouts with a train batch size of 128, and the current policy π_θ is updated

Table 2: Comparison of different RL algorithms trained on Qwen3-1.7B and Qwen3-4B across various mathematical reasoning benchmarks (evaluated by Avg@8). We highlight each cell as best, second best, and third best.

Method	Granularity	AIME24	AIME25	BRUMO25	SMT25	Average
Qwen3-1.7B-Base						
Base	none	2.1	0.0	11.3	6.6	5.0 (-55.0%)
GRPO	token-level	11.3	10.8	15.0	7.3	11.1 (+0.0%)
DAPO	token-level	15.0	12.1	9.2	7.8	11.0 (-0.7%)
DCPO	token-level	17.1	11.3	15.4	10.8	13.7 (+23.0%)
RIPO	token-level	18.3	12.9	15.4	10.4	14.3 (+28.4%)
GSPO	sequence-level	16.3	11.2	12.9	8.5	12.2 (+10.1%)
GMPO	sequence-level	17.5	12.5	15.0	6.6	12.9 (+16.2%)
TAPO	trajectory-aware	18.9	12.5	16.2	10.8	14.6 (+31.5%)
Qwen3-4B-Base						
Base	none	10.0	8.3	18.3	8.9	11.4 (-53.0%)
GRPO	token-level	30.4	20.8	25.8	19.8	24.2 (+0.0%)
DAPO	token-level	31.7	21.3	26.3	21.7	25.3 (+4.3%)
DCPO	token-level	31.7	24.2	25.0	19.8	25.2 (+4.0%)
RIPO	token-level	32.9	27.5	31.3	21.5	28.3 (+16.9%)
GSPO	sequence-level	32.1	21.3	26.7	18.2	24.6 (+1.5%)
GMPO	sequence-level	32.5	26.7	26.3	22.5	27.0 (+11.6%)
TAPO	trajectory-aware	32.9	28.7	33.3	22.5	29.5 (+21.3%)

8 times with a mini-batch size of 16. Rewards are set to 1 for correct responses and 0 otherwise. We uniformly adopt token-mean loss aggregation and optimize all models to convergence using AdamW (Loshchilov & Hutter, 2017) with a constant learning rate of 1×10^{-6}, training for 750 steps. All experiments are conducted on VeRL (Sheng et al., 2025) framework using 8×H20 GPUs.

Evaluation. We comprehensively evaluate the effectiveness of these algorithms on eight competition-level benchmarks without contamination, spanning the two domains. For math we use AIME24, AIME25, BRUMO25, and SMT25 (Balunović et al., 2025); for coding we use Codeforces, CodeContests, TACO, and APPS (Cui et al., 2025a). We evaluate the best checkpoint of each algorithm, repeat the evaluation set for 8 times and report avg@8 metric for results stability.

4.2 Main Results

Table 1 summarizes the performance of different RL methods across eight competition-level mathematical reasoning and algorithmic coding benchmarks. TAPO consistently achieves the best performance across both task domains. Specifically, compared to GRPO, TAPO improves the average performance by 48.0% on math and 11.8% on coding. Compared to the strongest token-level correction method, RIPO, TAPO further improves the average performance by 1.2% and 2.4%, respectively. Moerover, compared to the sequence-level method GSPO, TAPO achieves improvements of 16.8% and 9.1% on math and coding.

Moreover, Fig. 2 shows the training dynamics of GRPO, GSPO, and TAPO on mathematical reasoning. As observed, GRPO and GSPO eventually suffer from training collapse, characterized by declining training scores and fluctuating entropy as the policy shift grows (reflected by the increasing KL divergence). In contrast, TAPO maintains steadily improving training scores with controlled KL divergence and stable entropy throughout training, validating that trajectory-aware correction effectively mitigates the state distribution shift induced by long-horizon policy updates, leading to more stable and sustained training.

Overall, the experimental results persuasively demonstrate the effectiveness and generality of trajectory-aware policy optimization.

4.3 Generalization Across Model Scales

We further generalize TAPO to different model scales, including Qwen3-1.7B-Base and Qwen3-4B-Base, trained on mathematical reasoning. As shown in Table 2, TAPO consistently achieves the best average performance on both models, improving over GRPO by 31.5% and 21.3%, respectively.

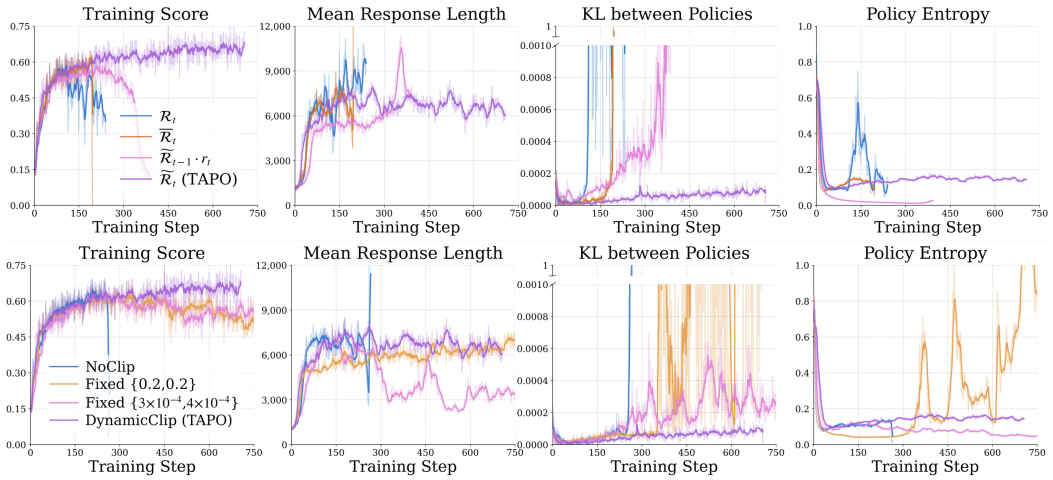


Figure 3: Training dynamics of different ablation studies on mathematical reasoning tasks. **Upper**: trajectory correction ratio variants. **Lower**: clipping strategies.

Compared with GMPO, TAPO further improves the average performance by 13.2% on Qwen3-1.7B and 9.3% on Qwen3-4B. These results demonstrate that the general effectiveness of trajectory-aware correction across model scales.

4.4 Ablation Studies

We further conduct ablation studies on Qwen3-8B-Base trained for mathematical reasoning to investigate the design choices of TAPO. Specifically, we examine two components: the construction of the trajectory-aware correction ratio and the length-aware clipping strategy.

Trajectory-aware Correction Ratio. For the trajectory correction ratio, we compare four variants: the original trajectory ratio $\mathcal{R}_t$; $\mathcal{R}_t$ with excessive samples outside the range $[e^{-0.4}, e^{0.4}]$ removed, following GMPO (denoted as $\overline{\mathcal{R}}_t$); geometric averaging applied only to the state correction ratio, resulting in $\widetilde{\mathcal{R}}_{t-1} \cdot r_t$; and geometric averaging applied to the full trajectory ratio $\mathcal{R}_t$, corresponding to the proposed $\widetilde{\mathcal{R}}_t$. As shown in the upper panel of Fig. 3, the original trajectory ratio $\mathcal{R}_t$ exhibits highly unstable training dynamics due to the unbounded variance introduced by multiplying correction ratios over the trajectory. Removing excessively large correction ratios following GMPO does not resolve this issue and eventually still leads to training collapse. Applying geometric averaging only to the state correction ratio, i.e., $\widetilde{\mathcal{R}}_{t-1} \cdot r_t$, essentially add an additional scaling factor to the action-level correction, making the clipping range difficult to calibrate and ultimately ineffective. In contrast, geometric averaging applied to the full trajectory ratio yields stable training dynamics and sustained score improvement. These results demonstrate that controlling the variance of the complete trajectory correction is essential for effective trajectory-aware policy optimization.

Clipping Startegy. We further compare four clipping strategies: no clipping; fixed clipping with $\epsilon_t^l = \epsilon_t^h = 0.2$ (as adopted by GRPO); fixed clipping with $\epsilon_t^l = 0.0003, \epsilon_t^h = 0.0004$ (as adopted by GSPO); and the proposed length-aware dynamic clipping. As shown in the lower panel of Fig. 3, removing the proximal constraint of clipping leads to sudden training collapse. GRPO-order clipping boundary provides insufficient constraint for later tokens, resulting in increasingly unstable KL divergence and entropy. In contrast, GSPO-order clipping boundary imposes overly restrictive constraints on earlier tokens, causing entropy to decrease rapidly and slowing score improvement. These results demonstrate the importance of adapting the clipping boundary to trajectory length.

5 Conclusion

This paper introduces Trajectory-Aware Policy Optimization (TAPO), a principled framework for correcting the policy distribution shift in long-horizon LLM RL, derived from the off-policy policy gradient. Across eight mathematical reasoning and algorithmic coding benchmarks, TAPO substantially outperforms existing RL algorithms, demonstrating the effectiveness of trajectory-aware policy correction. Ultimately, TAPO provides a principled foundation for advancing long-horizon autoregressive policy optimization.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading and grammatical refinement of the manuscript. We have not used generative AI tools for research ideation, theoretical derivation, experimental design, or code development, and automated decision-making or synthetic data generation tasks are not applicable to this work. Additionally, we used generative AI tools for minor stylistic and readability enhancements. We have reviewed all AI-assisted work. Specifically, all AI-suggested text modifications were manually reviewed and verified by the authors to ensure that the original research meaning and technical accuracy remained intact. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, the full source code and instructions for reproducing all experimental results are included in the supplementary material. The codebase includes environment setup files, training and evaluation scripts, and default hyperparameters. We will also open-source the codebase upon publication.

REFERENCES

Anthropic. Claude for coding. `https://www.anthropic.com/claude`, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmark*, 2025.

Zhicheng Cai, Xinyuan Guo, Hanlin Wu, Mingxuan Wang, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Beyond euclidean clipping: Overcoming exploration collapse in llm rl via riemannian isometric policy optimization. *International conference on machine learning*, 2026.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

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*, 2025b.

Shihan Dou, Yan Liu, Haoxiang Jia, Limao Xiong, Enyu Zhou, Wei Shen, Junjie Shan, Caishuang Huang, Xiao Wang, Xiaoran Fan, et al. Stepcoder: Improve code generation with reinforcement learning from compiler feedback. *arXiv preprint arXiv:2402.01391*, 2024.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *Proceedings of the Conference on Language Modeling (COLM)*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *International Conference on Learning Representations*, 2025.

OpenAI. Learning to reason with large language models. https://openai.com/research/learning-to-reason-with-llms, 2024.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *International conference on machine learning*, 2026.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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.

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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6672–6679, 2020.

Qiying 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. *Proceedings of the Neural Information Processing Systems*, 2025.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, et al. Geometric-mean policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 41354–41374, 2026.

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.