arXiv:2602.04879v1 [cs.LG] 4 Feb 2026

Rethinking the Trust Region in LLM Reinforcement Learning

Penghui Qi [*] [1] [2] **Xiangxin Zhou** [*] [1] **Zichen Liu** [2] **Tianyu Pang** [1] **Chao Du** [1] **Min Lin** [1] **Wee Sun Lee** [2]

https://github.com/sail-sg/Stable-RL

$$L_\mu(\pi) = \mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{|y|} M_t \cdot \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} \cdot \hat{A}_t\right],$$

$$M_t^{\text{PPO}} = \begin{cases} 0, & \text{if } \hat{A}_t > 0 \text{ and } \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} > 1 + \epsilon_{\text{high}}, \\ 0, & \text{if } \hat{A}_t < 0 \text{ and } \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} < 1 - \epsilon_{\text{low}}, \\ 1, & \text{otherwise}, \end{cases}$$

$$M_t^{\text{DPPO}} = \begin{cases} 0, & \text{if } \hat{A}_t > 0 \text{ and } \pi(y_t|s_t) - \mu(y_t|s_t) > \delta, \\ 0, & \text{if } \hat{A}_t < 0 \text{ and } \mu(y_t|s_t) - \pi(y_t|s_t) > \delta, \\ 1, & \text{otherwise}. \end{cases}$$

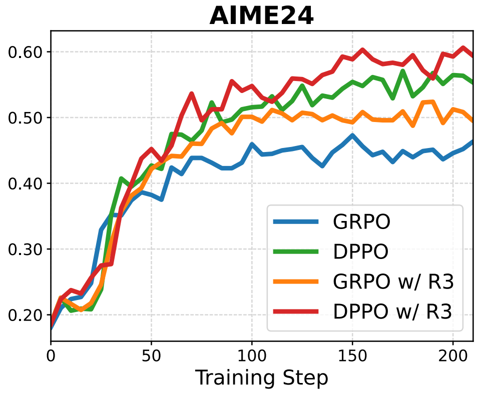

Figure 1. Comparison of PPO and the proposed DPPO (the Binary-TV variant in Section 4.4). (**Left**) The surrogate objective and corresponding masks for PPO and DPPO. PPO (and variants like GRPO) employs a heuristic mask based on the probability ratio, which **over-penalizes low-probability tokens** and **under-penalizes high-probability ones** (Section 4.2). In contrast, DPPO utilizes a more principled mask based on a direct approximation of policy divergence (e.g., Total Variation), ensuring updates stay within a theoretically grounded trust region (Section 3). (**Right**) Experimental results on the AIME24 using Qwen3-30B-A3B-Base. DPPO significantly outperforms GRPO baselines, achieving superior training efficiency and stability even without rollout routing replay (R3) (Section 7).

Abstract

Reinforcement learning (RL) has become a cornerstone for fine-tuning Large Language Models (LLMs), with Proximal Policy Optimization (PPO) serving as the de facto standard algorithm. Despite its ubiquity, we argue that the core ratio clipping mechanism in PPO is structurally ill-suited for the large vocabularies inherent to LLMs. PPO constrains policy updates based on the probability ratio of sampled tokens, which serves as a noisy single-sample Monte Carlo estimate of the true policy divergence. This creates a sub-optimal learning dynamic: updates to low-probability tokens are aggressively over-penalized, while potentially catastrophic shifts in high-probability tokens are under-constrained, leading to training inefficiency and instability. To address this, we propose **Divergence Proximal Policy Optimization (DPPO)**, which substitutes heuristic clipping with a more principled constraint based on a direct estimate of policy divergence (e.g., Total Variation or KL). To avoid huge memory footprint, we introduce the efficient Binary and Top-K approximations to capture the essential divergence with negligible overhead. Extensive empirical evaluations demonstrate that DPPO achieves superior training **stability** and **efficiency** compared to existing methods, offering a more robust foundation for RL-based LLM fine-tuning.

1. Introduction

Reinforcement learning (RL) is a foundational paradigm for fine-tuning Large Language Models (LLMs), enabling alignment with human preferences (Ouyang et al., 2022; Rafailov et al., 2023) and complex reasoning tasks (Guo et al., 2025; Qi et al., 2025a). Proximal Policy Optimization (PPO[1]) (Schulman et al., 2017) has established itself as the de facto standard in this domain, favored for its simplicity and empirical scalability. Central to PPO is a heuristic clipping mechanism designed to prevent destructive policy updates. By constraining the probability ratio between new

[*]Equal contribution [1]Sea AI Lab, Singapore [2]School of Computing, National University of Singapore. Correspondence to: Wee Sun Lee <leews@comp.nus.edu.sg>, Penghui Qi <penghuiq@comp.nus.edu.sg>.

Preprint. February 5, 2026.

[1] We denote PPO by its ratio-clipping loss, regardless of advantage estimation. Under this definition, GRPO is a PPO variant.

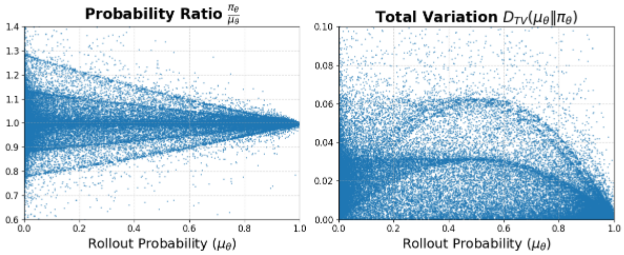

Figure 2. The plots show numerical differences between a training and an inference engine for Qwen3-30B-A3B-Base with identical parameters. **(Left)** The probability ratio (used in PPO) is highly volatile for low-probability tokens. **(Right)** In contrast, the TV divergence (used in DPPO) is more stable. This highlights a key flaw of PPO's clipping mechanism: it over-penalizes low-probability tokens, which can slow down learning; and under-penalizes high-probability tokens, which can permit large, destabilizing updates.

and old policies, the algorithm aims to confine learning to a *trust region* where monotonic improvement is theoretically guaranteed (Schulman et al., 2015).

Despite its widespread adoption, we argue that PPO's core mechanism, ratio clipping, is structurally ill-suited for the expansive, long-tailed vocabularies inherent to LLMs. Unlike Trust Region Policy Optimization (TRPO) (Schulman et al., 2015), which constrains the KL or Total Variation (TV) divergence of the policy distribution, PPO constrains updates based on the probability ratio of the sampled token. As we demonstrate later, this approach functions as a noisy, single-sample Monte Carlo estimate of the true policy divergence. While this approximation suffices for classical RL environments with limited action spaces, it fails in the LLM regime due to the ratio's hypersensitivity to the probability magnitude. For example, increasing a rare token's probability from 10^{-5} to 10^{-3} generates a massive ratio of 100 that triggers clipping, even though the actual divergence is negligible. Conversely, small ratio changes on high-probability tokens can make catastrophic shifts in probability mass (e.g., a drop from 0.99 to 0.8), yet it often remains unpenalized by the clipping mechanism.

This implicit bias is exacerbated by the *training-inference mismatch* (Yao et al., 2025; Qi et al., 2025b; Zheng et al., 2025), where numerical discrepancies arise between training and inference engines even under identical parameters. As illustrated in Figure 2, the probability ratio becomes highly volatile for low-probability tokens, while TV divergence remains stable. Consequently, PPO creates a sub-optimal learning dynamic: updates to low-probability tokens are aggressively over-penalized, slowing learning, while updates to high-probability tokens are under-penalized, risking instability. These limitations necessitate a fundamental rethinking of the trust region approach in LLM fine-tuning to ensure both efficiency and stability.

To address these fundamental limitations, we propose Divergence Proximal Policy Optimization (DPPO), a framework that substitutes PPO's heuristic clipping with a more principled constraint grounded in trust region theory. Rather than relying on noisy single-sample ratios, DPPO directly estimates policy divergence (e.g., TV or KL divergence). To ensure memory feasibility for LLMs, we introduce two efficient approximations, Binary and Top-K divergence, which capture essential distributional shifts with negligible overhead. This allows DPPO to rigorously distinguish between safe and unsafe updates, effectively resolving the problems of over- and under-constraining inherent in standard PPO.

In this work, we provide a comprehensive rethinking of the trust region in the context of LLM fine-tuning. Our contributions are threefold. **Theoretical Formulation**: We derive policy improvement bounds specifically tailored to the finite-horizon, undiscounted setting of LLM generation, establishing a rigorous theoretical foundation for trust-region methods in this domain. **Stability and Efficiency Analysis**: We isolate the primary sources of training instability to provide practical stabilization guidelines, while further highlighting the significant role that low-probability tokens play in driving exploration. **Algorithmic Performance**: We demonstrate that DPPO achieves superior stability and final performance compared to existing methods like GRPO, providing a robust new framework for RL-based fine-tuning.

2. Background

2.1. Policy Performance Difference

We begin with the standard formulation of a Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \rho_0, \gamma)$, which includes the state space $\mathcal{S}$, action space $\mathcal{A}$, transition dynamics $P(s'|s,a)$, reward function $r(s,a)$, initial state distribution $\rho_0(s)$, and a discount factor $\gamma \in [0,1]$. A stochastic policy $\pi(a|s)$ generates trajectories $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ by sampling actions $a_t \sim \pi(\cdot|s_t)$ and transitioning to states $s_{t+1} \sim P(\cdot|s_t, a_t)$. The central goal of RL is to find a policy that maximizes the expected discounted return:

$$\eta(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right].$$

To facilitate policy optimization, we define the standard value functions under a policy π: the state-value function $V^\pi(s) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s\right]$, the action-value function $Q^\pi(s,a) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a\right]$, and the advantage function $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. A key theoretical tool for relating the performance of two distinct policies is the policy performance difference theorem (Kakade & Langford, 2002). It states that for any two policies, a target policy (to be optimized) π and a behavior policy (for rollout) μ, their expected returns are related by:

$$\eta(\pi) - \eta(\mu) = \frac{1}{1-\gamma}\,\mathbb{E}_{s\sim\rho^{\pi},\, a\sim\pi(\cdot|s)}\big[A^{\mu}(s,a)\big]. \tag{1}$$

Here, $\rho^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^{t}\Pr(s_t = s|\,\pi)$ is the normalized discounted state-visitation distribution induced by the policy π. This identity is fundamental, as it implies that any policy update that results in a non-negative expected advantage guarantees monotonic performance improvement, i.e., $\eta(\pi) \geq \eta(\mu)$.

2.2. Policy Improvement Bound

While Equation 1 provides a direct expression for policy improvement, its dependence on the state-visitation distribution ρ^{π} of the new policy makes it intractable for direct optimization. To overcome this, Schulman et al. (2015) derive a lower bound on performance improvement that can be estimated using samples from the behavior policy μ, with a penalty term that measures the divergence between the old and new policies. This lower bound forms the basis of trust-region methods.

Theorem 2.1. *(Schulman et al., 2015; Achiam et al., 2017) Given any two policies, μ and π, the following bound holds:*

$$\begin{aligned}\eta(\pi)-\eta(\mu) \geq & \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho^{\mu},\, a\sim\mu(\cdot|s)}\left[\frac{\pi(a|s)}{\mu(a|s)}A^{\mu}(s,a)\right] \\ & - \frac{2\xi\gamma}{(1-\gamma)^2}D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi)^2,\end{aligned} \tag{2}$$

where $\xi = \max_{s,a}\left|A^{\mu}(s,a)\right|$ *and* $D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi) = \max_s D_{\mathrm{TV}}\big(\mu(\cdot|s)\|\,\pi(\cdot|s)\big)$*, which is the maximum Total Variation (TV) divergence among all states.*

This bound provides a direct path to guaranteed policy improvement. The right-hand side of the inequality forms a surrogate objective that is a tight lower bound on the true performance improvement, touching the objective when $\pi = \mu$. Therefore, iteratively maximizing this surrogate guarantees monotonic improvement in $\eta(\pi)$, following the principles of the Minorize-Maximization (MM) algorithm (Hunter & Lange, 2004; Schulman et al., 2015).

2.3. Trust Region Policy Optimization

The policy improvement bound in Equation (2) directly justifies a surrogate objective,

$$L_{\mu}(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho^{\mu},\, a\sim\mu(a|s)}\left[\frac{\pi(a|s)}{\mu(a|s)}A^{\mu}(s,a)\right]. \tag{3}$$

This objective serves as a **first-order approximation** of the true performance improvement $\eta(\pi) - \eta(\mu)$, as their values and gradients match at the point of expansion $\pi = \mu$ (Kakade & Langford, 2002; Schulman et al., 2015; Zheng et al., 2025). Therefore, maximizing $L_{\mu}(\pi)$ within a small *trust region* guarantees stable and meaningful policy improvement. This insight motivates the trust-region optimization approach (Schulman et al., 2015; Xie et al., 2024), which involves maximizing $L_{\mu}(\pi)$ subject to a constraint that keeps the new policy π within a trust region around the current policy μ, thereby ensuring the validity of the approximation. Formally, this is expressed as the following constrained optimization problem:

$$\begin{aligned}\max_{\pi} \quad & L_{\mu}(\pi) \\ \text{s.t.} \quad & D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi) \leq \delta,\end{aligned} \tag{4}$$

where the constraint can also be applied on a KL divergence D_{KL}, justified via Pinsker's inequality:

$$D_{\mathrm{TV}}(\mu\,\|\,\pi)^2 \leq \tfrac{1}{2}D_{\mathrm{KL}}(\mu\,\|\,\pi).$$

3. Trust Region Under LLM Regime

In this section, we adapt the trust region framework to the specific context of LLM fine-tuning. This setting differs from the classical RL paradigm in two crucial ways. First, the learning problem is structured as an undiscounted ($\gamma = 1$) episodic task with a finite horizon T, which makes the original bound in Equation (2) ill-defined, as the $\frac{1}{1-\gamma}$ term diverges to infinity. Second, due to the sparse reward nature, advantages are often estimated at the sequence level (Shao et al., 2024), rather than on a per-token basis.

Formally, given a prompt x, a policy π (the LLM) generates a response $y = (y_1, \ldots, y_T)$ by sequentially sampling tokens. At each step t, the policy defines a conditional distribution $\pi(y_t|s_t)$ over the vocabulary $\mathcal{A}$, where the state $s_t = (x, y_1, \ldots, y_{t-1})$ consists of the prompt and previously generated tokens. The probability of the complete response is the product of these conditional probabilities: $\pi(y|x) = \prod_{t=1}^{T}\pi(y_t|s_t)$. After the full response is generated, a scalar reward $R(y,x)$ is provided. For brevity, we will omit the dependency on the initial prompt x and write the objective function as:

$$\mathcal{J}(\pi) = \mathbb{E}_{y\sim\pi}[R(y)].$$

We now derive performance difference identity and policy improvement bound tailored to this regime.

Theorem 3.1 (Performance Difference Identity for LLMs)**.** *In a finite-horizon setting (T) with no discount ($\gamma = 1$), for any two policies π and μ, the performance difference can be decomposed as:*

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = L'_{\mu}(\pi) - \Delta(\mu, \pi),$$

where $L'_{\mu}(\pi)$ is a surrogate objective defined as:

$$L'_{\mu}(\pi) = \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{|y|}\left(\frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} - 1\right)\right], \tag{5}$$

and $\Delta(\mu, \pi)$ is an error term given by:

$$\Delta(\mu,\pi) = \mathbb{E}_{y\sim\mu}\Bigg[R(y) \sum_{t=1}^{|y|}\left(\frac{\pi(y_t|s_t)}{\mu(y_t|s_t)}-1\right)\left(1-\prod_{j=t+1}^{T}\frac{\pi(y_j|s_j)}{\mu(y_j|s_j)}\right)\Bigg]. \tag{6}$$

This theorem provides an exact expression for the policy improvement. The surrogate $L'_\mu(\pi)$ represents a first-order approximation, while the error term Δ captures the higher-order effects of the policy change. To make this practical for optimization, we bound the error term.

Theorem 3.2 (Policy Improvement Bound for LLMs)**.** *In a finite-horizon setting (T) with no discount ($\gamma = 1$), the policy improvement is lower-bounded by:*

$$\mathcal{J}(\pi)-\mathcal{J}(\mu) \geq L'_\mu(\pi)-2\xi T(T-1)\cdot D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi)^2, \tag{7}$$

where $\xi = \max_y |R(y)|$ is the maximum absolute reward, and $D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi) = \max_{s_t} D_{\mathrm{TV}}\big(\mu(\cdot|s_t)\|\,\pi(\cdot|s_t)\big)$ is the maximum Total Variation (TV) divergence over all states.

This theorem establishes a lower bound on policy improvement, and it is structurally analogous to the bound in Theorem 2.1 (see Appendix B.4), with the horizon T playing a role similar to the effective horizon $\frac{1}{1-\gamma}$ in the discounted setting. It provides a clear theoretical justification for adapting the trust region approach into LLM regime. Similar to Equation (4), we can solve the following constrained optimization problem to guarantee stable learning:

$$\begin{aligned} \max_{\pi} \quad & L'_\mu(\pi) \\ \text{s.t.} \quad & D_{\mathrm{TV}}^{\max}(\mu\,\|\,\pi) \leq \delta, \end{aligned} \tag{8}$$

where the constraint can also be applied on a KL divergence.

The proofs for Theorem 3.1 and Theorem 3.2 are deferred to Appendix B. In Appendix B.3, we further derive a more practical bound that depends linearly, rather than quadratically, on the horizon length T.

4. Methodology

4.1. Proximal Policy Optimization

While theoretically appealing, the constrained optimization in TRPO requires second-order methods that are computationally expensive and difficult to scale. PPO (Schulman et al., 2017) was introduced to achieve the stability of a trust region method using only first-order optimization. Its simplicity and strong empirical performance have established it as a standard algorithm for fine-tuning LLMs.

Instead of an explicit trust region constraint, PPO discourages overly large policy updates with a heuristic clipping mechanism applied to the surrogate objective. Specifically, PPO optimizes the following clipped objective:

$$L_\mu^{\mathrm{PPO}}(\pi) = \mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{|y|}\min\Big(r_t\hat{A}_t, \operatorname{clip}(r_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\right],$$
$$r_t = \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)}, \tag{9}$$

where $\hat{A}_t$ is an estimated advantage at timestep t. In LLM fine-tuning, the advantage is usually estimated by $\hat{A} = R(y) - \frac{1}{G}\sum_{i=1}^{G} R(y_i)$ (Shao et al., 2024; Liu et al., 2025d), where $\{y_i\}_{i=1}^{G}$ is a group of responses for the same prompt.

The connection between PPO's clipping and the formal trust region can be understood by examining the TV divergence:

$$D_{\mathrm{TV}}\big(\mu(\cdot|s_t)\|\,\pi(\cdot|s_t)\big) = \frac{1}{2}\mathbb{E}_{y_t\sim\mu}\left[\left|r_t-1\right|\right]. \tag{10}$$

From this perspective, PPO's clipping condition, $|r_t - 1| \leq \epsilon$, can be interpreted as constraining a **single-sample Monte Carlo estimate** of the expected value in Equation (10). In essence, PPO enforces its trust region not on the true TV divergence, but on a noisy, single-point estimation. As we will argue next, this crude approximation is the source of significant pathologies when applied to the large, long-tailed vocabulary distributions characteristic of LLMs.

4.2. Limitations of PPO Ratio Clipping

The key limitation of PPO is that whether an update is clipped depends heavily on the sampled token's probability, rather than the true TV divergence between $\mu(\cdot|s_t)$ and $\pi(\cdot|s_t)$. Concretely, consider a fixed state s and two tokens a_{low} and a_{high} with

$$\begin{aligned} \mu(a_{\mathrm{low}}|s) &= 10^{-4}, \qquad & \pi(a_{\mathrm{low}}|s) &= 10^{-2}, \\ \mu(a_{\mathrm{high}}|s) &= 0.99, \qquad & \pi(a_{\mathrm{high}}|s) &= 0.80. \end{aligned}$$

The probability ratio for the low-probability token is $r_{\mathrm{low}} = \frac{10^{-2}}{10^{-4}} = 100$, which is far outside a typical clipping range $[1-\epsilon, 1+\epsilon]$ (e.g., $\epsilon = 0.2$). PPO would thus heavily clip the contribution of this update. In contrast, the actual contribution of this change to the TV divergence can be very small, because the total mass moved at a_{low} is tiny. For the high-probability token, $r_{\mathrm{high}} = \frac{0.80}{0.99} \approx 0.808$, which can still lie *inside* the clipping range for a moderate ϵ. Yet this update removes 0.19 probability mass from the dominant token, and therefore induces a much larger contribution to D_{TV}.

These examples highlight a structural flaw in PPO's clipping heuristic. For **low-probability tokens**, an update that produces a large probability ratio is aggressively constrained, even when its impact on the TV divergence is negligible, thereby slowing training efficiency. Conversely, for **high-probability tokens**, an update producing a ratio close to one may go unpenalized, even when the absolute change in probability mass is large enough to cause a substantial TV divergence, which in turn risks training instability.

Connections to Existing Work The insight that PPO's ratio clipping disproportionately penalizes low-probability tokens aligns with several prior studies. For instance, methods like *Clip-Higher* (Yu et al., 2025) and CISPO (Chen et al., 2025) observe that important "exploration" or "reasoning" tokens often have low initial probabilities (see Appendix E). These tokens usually get high importance ratios during policy updates and are consequently clipped, hindering the learning process. However, the solutions proposed remain heuristic and problematic. *Clip-Higher* suggests manually increasing the upper clipping bound, while CISPO continues to apply the gradient even for large divergence, completely ignoring the trust region. While these methods correctly identify the symptom, they fail to address the root cause: the fundamental mismatch between the single-sample probability ratio and the true distributional divergence.

4.3. Divergence Proximal Policy Optimization

To address the limitations of ratio clipping, we introduce Divergence Proximal Policy Optimization (DPPO), a method that replaces PPO's flawed heuristic with a more principled constraint grounded in trust region theory. Similar to Chen et al. (2025); Zheng et al. (2025), DPPO employs a dynamic mask to prevent updates that would leave the trust region. The DPPO objective is:

$$L_{\mu}^{\mathrm{DPPO}}(\pi) = \mathbb{E}_{y\sim\mu}\left[\sum_{t=1}^{|y|} M_t^{\mathrm{DPPO}} \cdot r_t \cdot \hat{A}_t\right]. \qquad (11)$$

Our key innovation lies in the design of this mask. Instead of relying on the noisy single-sample ratio, it is conditioned on a direct measure of the policy distribution's divergence:

$$M_t^{\mathrm{DPPO}} = \begin{cases} 0, & \text{if } (\hat{A}_t > 0 \text{ and } r_t > 1 \text{ and } D > \delta) \text{ or} \\ & \quad (\hat{A}_t < 0 \text{ and } r_t < 1 \text{ and } D > \delta) \\ 1, & \text{otherwise,} \end{cases} \qquad (12)$$

where $D \equiv D\big(\mu(\cdot|s_t) \| \pi(\cdot|s_t)\big)$ is the divergence (e.g., TV or KL) between the policy distributions, and δ is a divergence threshold hyperparameter.

This design directly approximates the formal trust region constraint from Theorem 3.2 while preserving the beneficial asymmetric structure of PPO's clipping. The mask only considers blocking an update if it is already moving away from the trusted region (i.e., $r_t > 1$ for a positive advantage or $r_t < 1$ for a negative advantage). It never blocks updates that move the policy ratio towards one (e.g., when $\hat{A}_t > 0$ and $r_t < 1$), a desirable property for accelerating learning.

Unlike PPO, the final decision to block an update is based on whether the entire policy distribution has shifted too far ($D > \delta$), not on the noisy and often misleading ratio of a single sample. This resolves the over- and under-constraining issues inherent in standard PPO. The primary remaining challenge is the overhead of calculating the full divergence D over a large vocabulary in LLMs, which we address next.

4.4. Approximating Distribution Divergence

Directly computing the policy divergence is memory-prohibitive for LLMs. To make it practical, we introduce two lightweight approximations, which serve as principled lower bounds of the true divergence (see Appendix C).

Binary Approximation The binary approximation collapses the original categorical distribution into a simple Bernoulli distribution, distinguishing only between the sampled token and all other tokens. We define the new distribution as: $p_t^{\tilde{\pi}} = \big(\tilde{\pi}(a_t|s_t), \quad 1 - \tilde{\pi}(a_t|s_t)\big)$, where $\tilde{\pi}$ can be μ or π. The TV and KL divergences are then computed as:

$$D_{\mathrm{TV}}^{\mathrm{Bin}}(t) = \big|\, \mu(a_t|s_t) - \pi(a_t|s_t)\big|, \qquad (13)$$

$$\begin{aligned} D_{\mathrm{KL}}^{\mathrm{Bin}}(t) &= \mu(a_t|s_t) \log \frac{\mu(a_t|s_t)}{\pi(a_t|s_t)} \\ &\quad + (1 - \mu(a_t|s_t)) \log \frac{1 - \mu(a_t|s_t)}{1 - \pi(a_t|s_t)}. \end{aligned} \qquad (14)$$

This binary divergence can be computed at negligible overhead. Crucially, it correctly distinguishes between large versus small shifts in absolute probability mass, thereby resolving the primary failure mode of PPO's clipping.

Top-K Approximation To provide a richer and more faithful approximation of the distributional shift, the top-K variant explicitly tracks the most probable tokens. First, we define a small, representative set of tokens $\mathcal{A}_t'$ as: $\mathcal{A}_t' = \mathrm{TopK}\big(\mu(\cdot|s_t), K\big) \cup \{a_t\}$, which includes the K highest-probability tokens under the behavior policy, augmented with the sampled token a_t if it is not already present. We then form reduced categorical distributions, p_t^{μ} and p_t^{π}, over the new vocabulary $\mathcal{A}_t'' = \mathcal{A}_t' \cup \{\text{other}\}$. For any token $a \in \mathcal{A}_t'$, its probability is its original probability, while all other tokens are aggregated into the "other" category: $p_t^{\tilde{\pi}}(a) = \tilde{\pi}(a|s_t) \quad \forall a \in \mathcal{A}_t'$, and $p_t^{\tilde{\pi}}(\text{other}) = 1 - \sum_{a\in\mathcal{A}_t'} \tilde{\pi}(a|s_t)$, where $\tilde{\pi}$ can be μ or π. The divergence is then computed over this reduced distribution:

$$D_{\mathrm{TV}}^{\mathrm{TopK}}(t) = \frac{1}{2} \sum_{a\in\mathcal{A}_t''} \big|p_t^{\mu}(a) - p_t^{\pi}(a)\big|, \qquad (15)$$

$$D_{\mathrm{KL}}^{\mathrm{TopK}}(t) = \sum_{a\in\mathcal{A}_t''} p_t^{\mu}(a) \log \frac{p_t^{\mu}(a)}{p_t^{\pi}(a)}. \qquad (16)$$

This approach better captures changes in the head of the policy distribution, which typically dominates the true divergence value. The overhead is minimal, making it a practical and high-fidelity choice for DPPO.

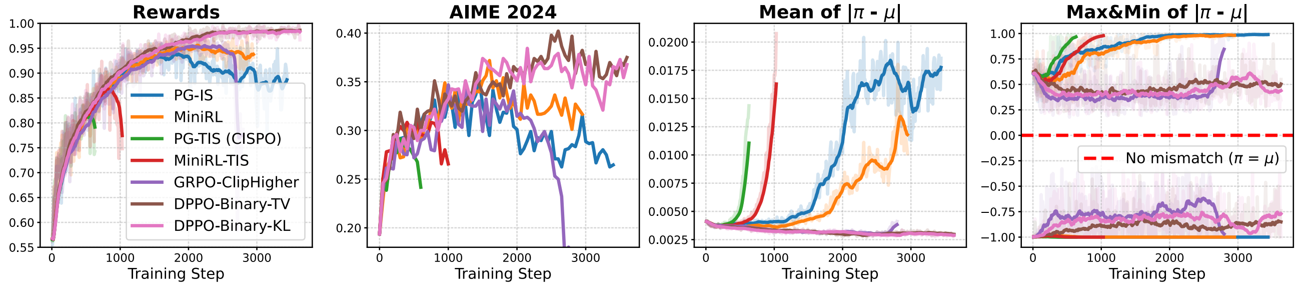

Figure 3. DPPO variants achieve stable training while controlling the training-inference mismatch at a low level. In contrast, methods without a trust region (PG-IS, CISPO) or with a misspecified one (MiniRL) suffer from growing mismatch and eventual collapse.

5. Analysis on Training Stability

The RL fine-tuning of LLMs is prone to training instability due to *training-inference mismatch* (see Appendix A.2). In this section, we conduct an empirical study to dissect this issue and verify the stability of our DPPO algorithm. To formalize our analysis, we denote the parameters being optimized as θ and the parameters used for data generation as θ'. We aim to answer three fundamental research questions:

1. Given the extremely low learning rates (e.g., 10^{-6}) common in LLM fine-tuning, is a trust region still necessary to ensure training stability?
2. Should the trust region be defined with respect to the original rollout distribution ($\mu_{\theta'}$) or a recomputed policy distribution ($\pi_{\theta'}$)?
3. What specific types of policy updates are the primary drivers of training instability?

Experimental Setting: Our experimental setup follows the sanity test proposed by Qi et al. (2025b). We fine-tune DeepSeek-R1-Distill-Qwen-1.5B (Guo et al., 2025) on a curated set of 1,460 problems from the MATH dataset (Hendrycks et al., 2021). In this setting, a stable algorithm should theoretically converge to 100% training accuracy, as all problems are known to be solvable by the initial model.

We evaluate several algorithms, each representing a different approach to managing the policy update. The baselines include: **PG-IS** and its truncated variant **PG-TIS** (also known as CISPO (Chen et al., 2025)), which use standard policy gradients with token-level importance sampling; **GRPO with Clip-Higher**, a PPO-like algorithm where clipping is based on the rollout policy ratio $r_t = \frac{\pi_\theta}{\mu_{\theta'}}$ (Shao et al., 2024; Liu et al., 2025d); and **MiniRL & MiniRL-TIS**, a PPO variant where clipping is based on a recomputed policy ratio $r_t = \frac{\pi_\theta}{\pi_{\theta'}}$ (Zheng et al., 2025). We compare these against **DPPO (Ours)**, our proposed method using either binary KL or TV divergence, with the trust region defined with respect to the rollout distribution $\mu_{\theta'}$. Detailed configurations for each algorithm are provided in the Appendix D.

5.1. The Necessity of a Trust Region

Our first question addresses whether a trust region is redundant at low learning rates. Figure 3 provides a clear answer. The unconstrained methods, PG-IS and PG-TIS (CISPO), both suffer from an increasing training-inference mismatch, which culminates in a collapse of performance. In contrast, our DPPO variants, which enforce a principled trust region, maintain a stable, low level of mismatch throughout training and achieve near-perfect final rewards.

Takeaway 1: A trust region is essential for stable training, even with very small learning rates. Without it, the training-inference mismatch accumulates and leads to collapse.

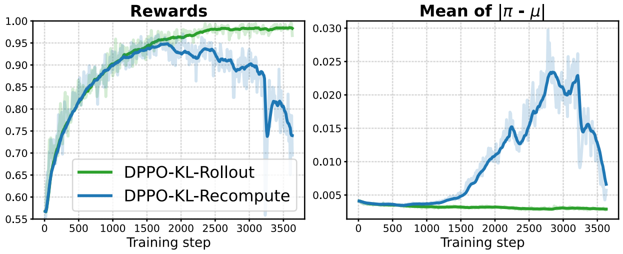

Figure 4. Switching the stable DPPO-KL to a decoupled objective causes the mismatch to grow and performance to collapse, confirming that the trust region must be anchored to the rollout policy.

5.2. The Correct Anchor for the Trust Region

Next, we investigate to which distribution the trust region should be anchored. A common practice in open-source implementations (Sheng et al., 2024; Zhu et al., 2025) is to use a *decoupled* objective (Hilton et al., 2022), where the trust region is enforced relative to a recomputed policy distribution ($\pi_{\theta'}$) instead of the original behavior policy ($\mu_{\theta'}$). The MiniRL algorithm, for example, follows this design (Zheng et al., 2025). Our results show this choice is detrimental. As in Figure 3, MiniRL fails to control the training-inference mismatch and its performance collapses, despite using a trust region. To confirm this, we created a decoupled version of our stable DPPO-KL algorithm. Figure 4 shows that this single change corrupts the stable training process, causing the mismatch to grow and performance to collapse.

Takeaway 2: The trust region must be defined with respect to the original behavior policy ($\mu_{\theta'}$). Using a recomputed on-policy distribution as the anchor leads to instability. This finding aligns with the theoretical bound in Equation (7) and offers a significant practical benefit: by removing the need for recomputation, we can reduce training costs by approximately 25% (Qi et al., 2025b).

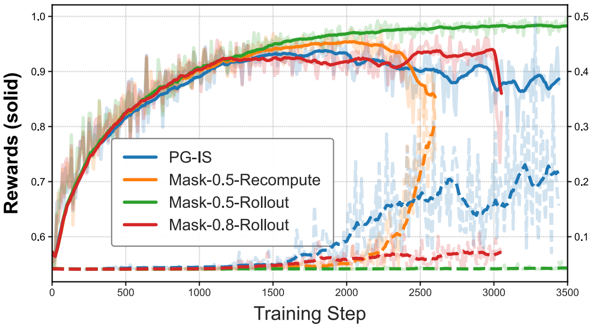

Figure 5. Isolating the source of instability. The solid curves are training rewards, while the dashed lines are the percentage of *bad updates*. Starting with the unstable PG-IS, applying a minimal mask that only blocks large-divergence bad updates on negative samples is sufficient to stabilize training, indicating these bad updates are the primary cause of training instability.

5.3. Identifying the Source of Instability

Finally, we seek to pinpoint which specific policy updates are most responsible for the instability. Our methodology is to start with the unstable PG-IS algorithm, which applies no update masking, and introduce the most minimal mask necessary to restore stability. This allows us to isolate the most detrimental class of updates. Since updates on positively rewarded samples are typically safe, we focus on negative samples where the policy is penalized (Liu et al., 2025a; Ren & Sutherland, 2025). We design a simple mask that only blocks updates on negative samples where the probability of the sampled token is decreased by more than a threshold δ: $M_t = 0$ if $\hat{A}_t < 0$ and $\mu_{\theta'}(y_t|s_t) - \pi_\theta(y_t|s_t) \geq \delta$. As shown in Figure 5, applying this minimal mask with $\delta = 0.5$ is sufficient to stabilize the training. In contrast, a slightly looser mask ($\delta = 0.8$) or one anchored to the recomputed distribution ("Mask-0.5-Recompute") both fail to prevent the eventual collapse. We define *bad updates* as those where this divergence exceeds 0.5 and plot their percentage over time. The plot reveals that only a very small fraction of updates are "bad" ($\leq 0.5\%$) yet they are the primary culprits behind training collapse. Furthermore, the percentage of these bad updates strongly correlates with reward fluctuation; as the fraction of bad updates rises, the reward curve becomes more erratic, reinforcing a causal link.

Takeaway 3: The primary source of instability is a small subset of updates on negative samples that push the policy far outside the trust region. A likely reason is that aggressively penalizing a token the model deems probable can corrupt the LLM's internal knowledge and destabilize the learning process. This finding confirms the critical need for a trust region, particularly when handling negative feedback.

5.4. The Pitfalls of Truncated Importance Sampling

Our empirical results also reveal a surprising finding regarding Truncated Importance Sampling (TIS), a technique widely adopted to control the variance of policy gradient estimates (Yao et al., 2025; Chen et al., 2025). Contrary to its intended purpose, TIS consistently degrades training stability in our experiments. As illustrated in Figure 3, the TIS-enabled variants (PG-TIS and MiniRL-TIS) suffer from premature collapse and significantly underperform their untruncated counterparts.

We hypothesize that this detrimental effect stems from the same issue as PPO's ratio clipping: low-probability tokens, which naturally produce high-variance ratios, are the most likely to be truncated by TIS. While this does reduce variance, it systematically down-weights the gradient signal from these tokens, introducing a significant and harmful bias into the policy update. This suggests that naive truncation can be just as damaging as naive clipping.

6. Analysis on Training Efficiency

Beyond training stability, the design of trust region is also critical for training *efficiency*. As motivated in Section 4.2, PPO's ratio-clipping over-constrains the updates to low-probability tokens, which might be permitted by a divergence-based trust region. In this section, we aim to analyze how low-probability tokens affect the training dynamics, thus justifying the adoption of divergence-based trust region in our DPPO algorithm.

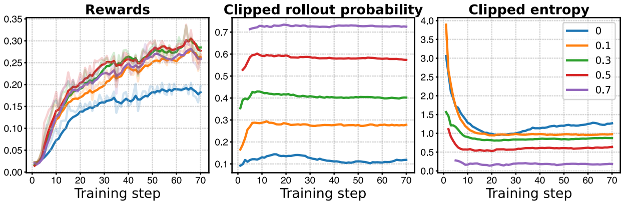

Figure 6. Analysis of relaxing trust regions for low-probability tokens. (**Left**) Training reward curves. (**Middle**) Rollout probability of clipped tokens. (**Right**) Entropy of clipped tokens.

Experimental Setting: We fine-tune Qwen3-1.7B-Base (Yang et al., 2025) on the DAPO dataset (Yu et al., 2025). We employ GRPO (Guo et al., 2025; Liu et al., 2025d) with the Clip-Higher trick (Yu et al., 2025) as the baseline algorithm. We then *relax* trust regions by setting the clipping threshold ϵ in Equation (9) as infinity for tokens with $\mu(y_t|s_t) < \alpha$, thus isolating the effect of low-probability tokens.

The learning curves for varying values of α are presented in Figure 6. Notably, relaxing the clipping constraint for tokens with $\mu(y_t|s_t) < 0.1$ yields a substantial improvement in training efficiency compared to the GRPO baseline ($\alpha = 0$). This observation validates our hypothesis that the ratio-clipping mechanism in PPO over-constrains updates to low-probability tokens, thereby hindering overall learning progress. The middle plot reveals that **clipped tokens are predominantly characterized by low probabilities** (typically below 0.15 for the baseline in blue). As α increases, the probabilities of clipped tokens also rise, confirming that PPO's ratio-clipping is structurally biased against low-probability tokens. Furthermore, the right plot demonstrates that **clipped tokens frequently exhibit high entropy**. Consistent with Wang et al. (2025a), which posits that RL is driven primarily by high-entropy tokens in LLMs, our results suggest that relaxing constraints on these tokens enables more informative policy updates and thus achieves higher training efficiency (see Appendix E for most frequent clipped tokens).

Furthermore, we examine the effect of directional clip relaxation with a fixed $\alpha = 0.1$. We generalize the clip operation with asymmetric thresholds, denoted as $\text{clip}(r_t, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})$, where $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$ by default. We relax either one end (*Relax-high* or *Relax-low*) or both ends (*Relax-both*). For example, Relax-high is implemented by $(\epsilon_{\text{low}} = 0.2, \epsilon_{\text{high}} = \infty)$ for tokens with $\mu(y_t|s_t) < \alpha$.

As illustrated in Figure 7, the direction of clip relaxation plays a critical role in the training efficiency and stability. Relax-high can be viewed as an extreme variant of the Clip-Higher trick (Yu et al., 2025) applied only to low-probability tokens. While this approach maintains high entropy, it fails to yield significant gains in training efficiency. Conversely, Relax-low exhibits substantially faster initial learning[2]. However, this strategy eventually drops due to entropy collapse (Cui et al., 2025). Ultimately, we find that **Relax-both is the most effective strategy for achieving both efficient and stable training**, thereby validating the design of DPPO in relaxing both ends of the trust region.

7. Scaling Experiments

Experimental Setting: We conduct large-scale experiments to further validate our methods. We train on a filtered subset of DAPO-Math (Yu et al., 2025), containing approximately 13k samples. Five model configurations (different base models and training techniques) are evaluated: (1) **MoE Base**: Qwen3-30B-A3B-Base (Yang et al., 2025); (2) **MoE Base w/ R3**: Qwen3-30B-A3B-Base with rollout router replay (R3) (Ma et al., 2025); (3) **MoE Thinking**: Qwen3-30B-A3B; (4) **Dense Base**: Qwen3-8B-Base; (5) **MoE Base w/ LoRA**: Qwen3-30B-A3B-Base with LoRA (Hu et al., 2022). Baseline methods include **GRPO-ClipHigher**(Shao et al., 2024; Liu et al., 2025d; Yu et al., 2025) and **CISPO**(Chen et al., 2025; Khatri et al., 2025). All methods use the behavior policy ($\mu_{\theta'}$) instead of recomputed policy distribution ($\pi_{\theta'}$) to construct the trust region (i.e., for clipping or masking). We compare our proposed methods, **DPPO-Binary-KL** and **DDPO-Binary-TV**, against these baselines. More details are provided in Appendix F.

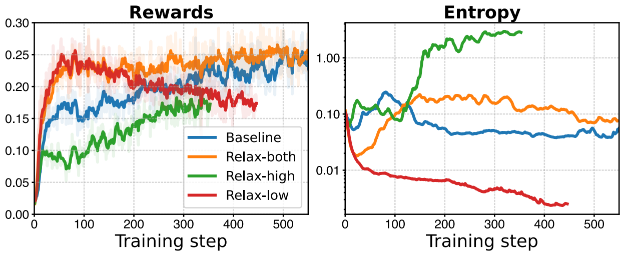

Figure 7. Analysis of trust region relaxation direction. (**Left**) Training reward curves. (**Right**) Policy entropy.

Main Results. We present online evaluation results on AIME24 and AIME25 (MAA, 2025) during RL training in the following figures: Figure 8 (MoE Base with and without R3) and Figure 9 (MoE Thinking and Dense Base). Results for MoE Base with LoRA are provided in Appendix G.1.

Our proposed method consistently demonstrates superior **stability** and **efficiency** across all five large-scale experiments. Specifically, DPPO optimizes rewards at a significantly faster speed than the GRPO-ClipHigher baseline and achieves better converged performance, providing empirical validation for the motivations discussed in Section 4.2. While all baseline methods frequently exhibit training instability or catastrophic collapse (e.g., CISPO in MoE Base without R3 and GRPO-ClipHigher in MoE Thinking), our approach maintains a remarkably stable training process.

Rollout router replay (R3) is widely considered a necessary technique for stabilizing RL training in MoE models (Ma et al., 2025; Zheng et al., 2025; Liu et al., 2025a). However, as illustrated in Figure 8, our DPPO variants (*without* R3) even consistently **outperform the R3-enhanced baselines**, which underscores the superior training efficiency and inherent stability of the DPPO framework. We provide additional detailed results and extended discussions in Appendix G.1.

Ablation on TV/KL Approximation. In the above scaling experiments, DPPO is implemented using the binary TV/KL

[2] In contrast to the Clip-Higher intuition (Yu et al., 2025), we observe that "Clip-Lower" (relaxing ϵ_{low}) for low-probability tokens is more vital for efficiency. This aligns with findings by Tajwar et al. (2024) regarding the role of negative gradients in accelerating preference learning.

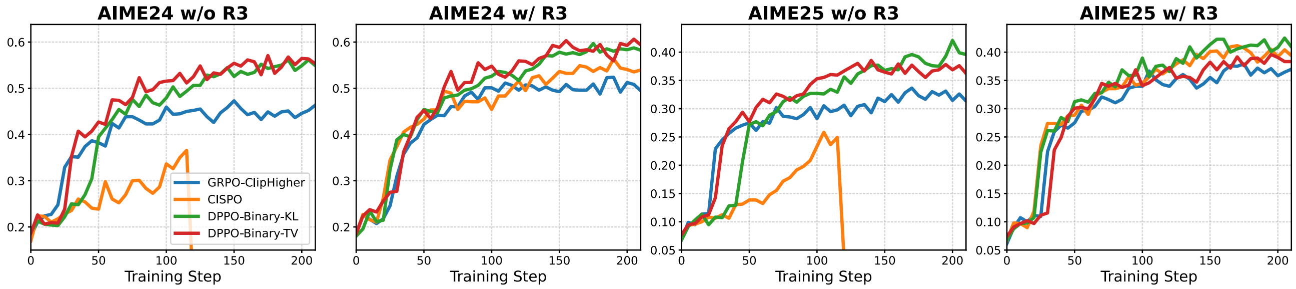

Figure 8. Evolution of AIME24 and AIME25 Avg@32 scores during RL training using Qwen3-30B-A3B-Base. The first and third panels correspond to the same experiment without rollout router replay (w/o R3), while the second and fourth panels correspond to the same experiment with rollout router replay (w/ R3).

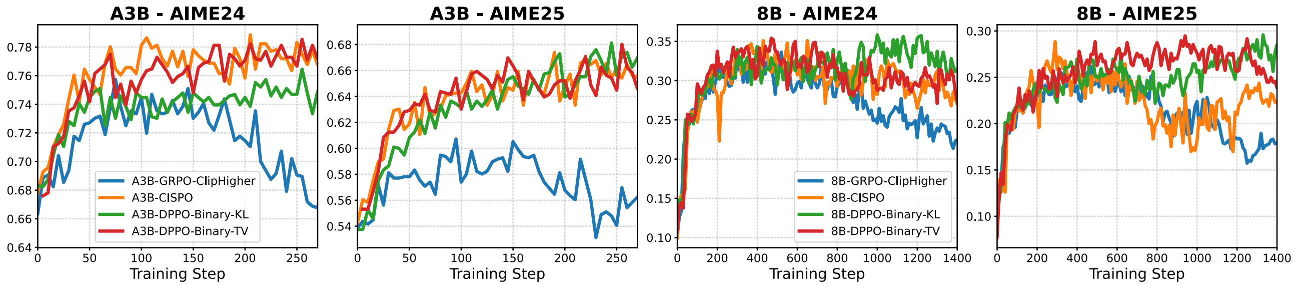

Figure 9. Evolution of AIME24 and AIME25 scores during RL training using Qwen3-30B-A3B (left) and Qwen3-8B-Base (right).

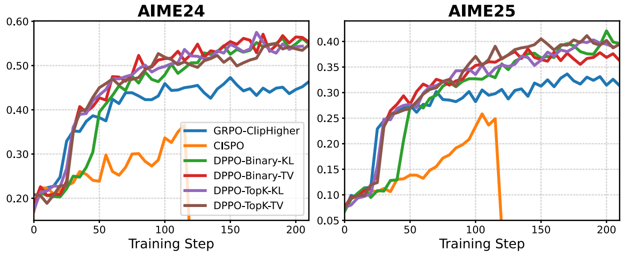

Figure 10. Evolution of AIME24 and AIME25 scores for baselines and DPPO with binary/Top-K (K=20) TV/KL approximation under the same setting as MoE Base w/o R3.

approximation (Equations 13 and 14). To assess the impact of this simplification, we compare it against DPPO with the top-K (K=20) TV/KL (Equations 15 and 16) under the same setting as MoE Base. The results, presented in Figure 10, show that both approximations perform similarly and significantly outperform the baselines. This finding indicates that the easy-to-implement binary approximation is a sufficient and computationally efficient choice for scalable RL. We provide more detailed results in Appendix G.2.

Generalization to Other Model Families and Tasks. We also conduct experiments on models from the Llama family (Touvron et al., 2023; Wang et al., 2025b) and on tasks beyond math reasoning (Liu et al., 2025e). The results, which are presented in Appendix G.3, show DPPO outperforms the baseline across most settings, highlighting its broad applicability.

8. Conclusion

In this work, we have presented a comprehensive rethinking of the trust region framework within the context of LLM fine-tuning. We derived policy improvement bounds specifically tailored to the finite-horizon, undiscounted setting of LLM generation, establishing a rigorous theoretical foundation for future trust-region research. Furthermore, through extensive empirical analysis, we investigated the trade-offs between training stability and efficiency, providing practical guidelines to optimize both.

Central to our contribution is the introduction of Divergence Proximal Policy Optimization (DPPO). We identified and addressed a critical structural flaw in the standard PPO algorithm: it over-constrains updates to low-probability tokens while under-constraining potentially catastrophic shifts in high-probability tokens. This implicit bias results in a suboptimal training dynamic, particularly for the expansive, long-tailed vocabularies inherent to LLMs. By substituting heuristic ratio clipping with a more principled policy divergence, DPPO significantly enhances both efficiency and stability. To avoid huge memory footprint for computing an exact policy divergence, we introduced Binary and Top-K approximations, which capture essential divergence with negligible overhead. Our evaluations demonstrate that DPPO consistently outperforms existing methods like GRPO in both training efficiency and stability, offering a more robust foundation for the RL-based LLM fine-tuning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *International conference on machine learning*, pp. 22–31. PMLR, 2017.

Chen, A., Li, A., Gong, B., Jiang, B., Fei, B., Yang, B., Shan, B., Yu, C., Wang, C., Zhu, C., et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Cui, G., Zhang, Y., Chen, J., Yuan, L., Wang, Z., Zuo, Y., Li, H., Fan, Y., Chen, H., Chen, W., et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

He, H. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hilton, J., Cobbe, K., and Schulman, J. Batch size-invariance for policy optimization. *Advances in Neural Information Processing Systems*, 35:17086–17098, 2022.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hunter, D. R. and Lange, K. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Khatri, D., Madaan, L., Tiwari, R., Bansal, R., Duvvuri, S. S., Zaheer, M., Dhillon, I. S., Brandfonbrener, D., and Agarwal, R. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025a.

Liu, J., Li, Y., Fu, Y., Wang, J., Liu, Q., and Shen, Y. When speed kills stability: Demystifying rl collapse from the inference-training mismatch, 2025b. https://yingru.notion.site/When-Speed-Kills-Stability-Demystifying-RL-Collapse-from-the-Inference-Training-Mismatch-271211a558b7808d8b12d403fd15edda.

Liu, Z., Chen, C., Du, C., Lee, W. S., and Lin, M. Oat: A research-friendly framework for llm online alignment. https://github.com/sail-sg/oat, 2025c.

Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025d.

Liu, Z., Sims, A., Duan, K., Chen, C., Yu, S., Zhou, X., Xu, H., Xiong, S., Liu, B., Tan, C., et al. Gem: A gym for agentic llms. *arXiv preprint arXiv:2510.01051*, 2025e.

Ma, W., Zhang, H., Zhao, L., Song, Y., Wang, Y., Sui, Z., and Luo, F. Stabilizing moe reinforcement learning by aligning training and inference routers. *arXiv preprint arXiv:2510.11370*, 2025.

MAA. American invitational mathematics examination - aime. https://maa.org/, 2025.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Qi, P., Liu, Z., Pang, T., Du, C., Lee, W. S., and Lin, M. Optimizing anytime reasoning via budget relative policy optimization. *arXiv preprint arXiv:2505.13438*, 2025a.

Qi, P., Liu, Z., Zhou, X., Pang, T., Du, C., Lee, W. S., and Lin, M. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025b.