# Multi-Step Likelihood-Ratio Correction for Reinforcement Learning with Verifiable Rewards

**Deokgyu Yoon**[1]* **Hyungkyu Kang**[2]* **Joongkyu Lee**[1]* **Byeongchan Kim**[1]
**Gyungin Shin**[2] **Sungrae Park**[2] **Min-hwan Oh**[1]

[1]Seoul National University [2]Upstage

{hhbdgy,jklee0717,bckim97,minoh}@snu.ac.kr
{hkang,noel,sungrae.park}@upstage.ai

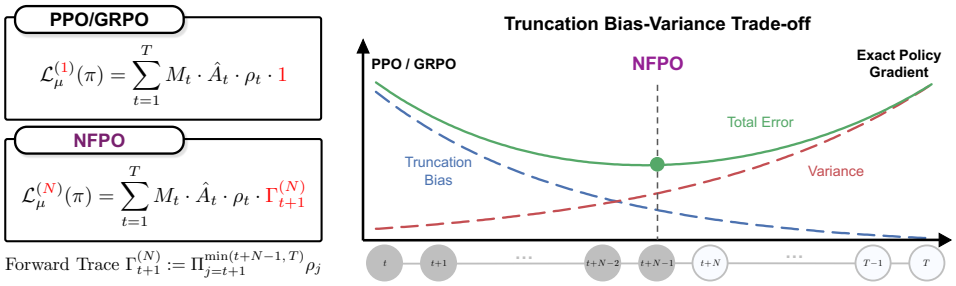



Figure 1: **NFPO** corrects the bias of *local* PPO/GRPO-style objectives by reweighting each token likelihood ratio $\rho_t := \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)}$ with a *forward trace* $\Gamma_{t+1}^{(N)}$. (Right) The trace horizon $N$ controls the bias–variance trade-off, interpolating between PPO/GRPO ($N = 1$) and the *exact* policy gradient objective ($N = T$).

## Abstract

Reinforcement learning with verifiable rewards (RLVR) plays a pivotal role in improving the reasoning ability of large language models. However, widely used PPO surrogate objectives are fundamentally *local*, as they rely on a local approximation of the exact policy gradient objective. While this approximation improves stability by reducing the variance induced by importance sampling, it also introduces structural bias into the surrogate objective, which must be controlled through trust region mechanisms. In this work, we introduce the $N$-step forward trace, which augments the PPO surrogate objective using the cumulative likelihood ratio of the next $N-1$ tokens. Building on this idea, we propose $N$-Step Forward-Trace Policy Optimization (NFPO), a practical RLVR algorithm that integrates the $N$-step forward trace into the masked policy gradient framework. NFPO provides a continuous bridge between the PPO surrogate objective and the exact policy gradient objective, offering a principled mechanism for controlling the bias–variance trade-off. Our theoretical analysis shows that, with an appropriate choice of $N$, the proposed objective yields a tighter policy-improvement bound than the standard PPO surrogate. Experiments on comprehensive reasoning benchmarks demonstrate that NFPO consistently improves performance, supporting our theoretical findings.

*Equal contribution. Code is available at `https://github.com/oh-lab/NFPO`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

# 1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a central paradigm for improving the reasoning ability of large language models (LLMs) [33, 9, 45]. By optimizing model-generated responses against verifier-provided rewards, *Proximal Policy Optimization* (PPO)-style policy optimization [32] has been widely adopted in recent RLVR methods, including Group Relative Policy Optimization (GRPO) [33, 9], yielding strong empirical gains on complex reasoning tasks.

Despite these successes, existing PPO-style surrogate objectives remain fundamentally *local* in how they account for policy change. Fix a prompt $x$. Let $y = (y_1, \ldots, y_T)$ denote a response, $\mu$ the rollout policy, and $\pi$ the target policy. The PPO surrogate objective optimizes a sum of token-level surrogate terms, each using the current-token likelihood ratio $\rho_t := \frac{\pi(y_t|x,y_{<t})}{\mu(y_t|x,y_{<t})}$ together with a response-level reward $R(y)$ or advantage. This locality makes training simple and stable, but it omits the effect of the updated policy on the probability of the remaining generated tokens.

This limitation becomes explicit through the exact policy-improvement identity (Lemma 1). For the expected reward $\mathcal{J}(\pi) := \mathbb{E}_{y\sim\pi}[R(y)]$, the following holds:

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\Gamma_{t+1}\right], \qquad \text{where} \quad \Gamma_{t+1} := \prod_{j=t+1}^{T}\rho_j.$$

Here, $\Gamma_{t+1}$ is the forward likelihood-ratio correction for the tokens generated after step $t$. Throughout the paper, we call $\Gamma_{t+1}$ the *full forward trace*.[2] The term "trace" emphasizes that this factor follows the generated trajectory forward from token $t+1$ to the end of the response, multiplying the likelihood-ratio corrections along the way.

The PPO surrogate objective corresponds to the case $\Gamma_{t+1} = 1$, yielding a biased approximation of the exact policy gradient objective. In particular, the omitted forward correction is governed by the deviation between the future trajectory distributions induced by $\mu$ and $\pi$: conditional on $s_{t+1}$, $\mathbb{E}_\mu[|\Gamma_{t+1} - 1| \mid s_{t+1}]$ is controlled by the total variation distance between these future distributions. Therefore, under the uniform trust-region condition $\sup_s D_{\mathrm{TV}}(\mu(\cdot|s), \pi(\cdot|s)) \leqslant \delta$, the resulting approximation error is bounded by $\mathcal{O}(T^2\delta^2)$ (Proposition 1). Thus, the trust region controls not only optimization stability, but also the structural bias induced by localizing the exact objective. In practice, PPO-based RLVR methods approximate this trust-region control through ratio-based clipping [32, 33, 9] or token-level update masking [38, 28].

However, trust-region mechanisms only indirectly compensate for the omitted full forward trace; they do not recover the missing forward likelihood-ratio correction itself. This leaves a structural bias–variance dilemma in policy improvement: the local surrogate is *stable* but *biased*, whereas the full forward trace is *unbiased* but can yield highly *unstable empirical estimates*, especially in long-horizon generation. Indeed, empirical optimization relies on a finite batch of rollouts, and the full trace $\Gamma_{t+1}$ can vary sharply across samples because it multiplies many token-level likelihood ratios. Consequently, a small number of trajectories with large forward traces can dominate the empirical gradient, leading to high variance and unstable updates.

This motivates an intermediate objective that recovers part of the missing forward trace while avoiding the empirical instability of the full trace. To address this dilemma, we introduce **N-step forward trace**. Rather than either discarding the forward trace correction entirely or using the full correction, we retain only the next $N-1$ future likelihood ratios after each token, i.e., $\Gamma_{t+1}^{(N)} := \prod_{j=t+1}^{\min\{t+N-1,T\}}\rho_j$. This construction interpolates between the PPO surrogate and the exact policy gradient objective: $N = 1$ gives $\Gamma_{t+1}^{(1)} = 1$, while $N = T$ recovers $\Gamma_{t+1}^{(T)} = \Gamma_{t+1}$. Our main theorem (Theorem 1) shows that increasing $N$ reduces the *truncation bias* at the rate $\mathcal{O}((T-N)^2\delta^2)$, but increases the *variance* exponentially in $N$, since longer forward traces involve products of more likelihood ratios. Thus, the trace horizon $N$ provides a principled knob for controlling the bias–variance trade-off in RLVR.

We further implement NFPO within a *masked* policy gradient framework [38, 28], and demonstrate through extensive reasoning benchmarks that NFPO consistently outperforms or matches both GRPO and DPPO. Beyond performance gains, we conduct a detailed mechanistic analysis of the forward

[2] This terminology is descriptive and should not be confused with *eligibility traces* in temporal-difference learning [36]. Eligibility traces are additive, recursively decayed variables for backward credit assignment, whereas our forward trace is a multiplicative likelihood-ratio correction over the future tokens $y_{t+1:T}$.

trace. Our analysis reveals that NFPO selectively concentrates forward trace correction on reasoning-related tokens, thereby improving the learning of reasoning behaviors. In addition, the forward trace acts as a semantic *low-pass filter* by aggregating learning signals from future tokens, which stabilizes the noisy token-level likelihood ratios that arise in local surrogate objectives. Finally, through experiments with varying trace horizons $N$, we empirically validate the bias–variance trade-off predicted by our main theorem. Together, these analyses complement our theoretical results by explaining how forward trace correction alters the underlying training dynamics.

Our main contributions are summarized as follows:

- **N-step bias–variance trade-off:** We revisit the standard PPO surrogate objective, and show that it discards the full forward trace in the exact policy-improvement identity, thereby inducing a structural bias (Proposition 1). We then prove that an $N$-step forward trace reduces *truncation bias* at the cost of higher *variance*, implying that an appropriately chosen trace horizon $N$ can yield a tighter policy-improvement lower bound than either the local surrogate (e.g., PPO) or the exact policy gradient (Theorem 1).
- **N-Step Forward-Trace Policy Optimization (NFPO):** Guided by the theory, we propose a practical policy-optimization algorithm that uses a truncated forward trace with an appropriate trace horizon $N$ (Section 3). The method bridges the stable but biased local PPO surrogate and the unbiased but high-variance exact policy gradient objective. On complex reasoning benchmarks, NFPO consistently outperforms baselines (Section 4).
- **Learning-dynamics analysis:** Our analyses show that forward trace amplifies token-level correction signals, concentrates updates on reasoning tokens, and reduces token-level learning noise. Experiments varying the trace horizon $N$ further highlight the importance of choosing an appropriate $N$, supporting the bias–variance trade-off predicted by our theory (Section 5).

## 2 Preliminaries

In this section, we introduce the basic setup and notation for policy optimization. Given a prompt dataset $\mathcal{D}$, a prompt $x \sim \mathcal{D}$ is sampled, and a large language model generates a response $y = (y_1, \ldots, y_T)$ autoregressively according to the policy $\pi(y|x) = \prod_{t=1}^{T} \pi(y_t|x, y_{<t})$. A verifier then assigns a scalar reward $R(x, y) \in [0, 1]$ to each prompt-response pair $(x, y)$. To simplify notation, throughout the paper we condition on a fixed prompt $x$, suppress the dependence on $x$, and write

$$\mathcal{J}(\pi|x) := \mathbb{E}_{y\sim\pi(\cdot|x)}[R(x, y)] \equiv \mathcal{J}(\pi) := \mathbb{E}_{y\sim\pi}[R(y)].$$

Let $\mu$ denote the behavior policy (for rollout) and $\pi$ the target policy (to be optimized). We define the token-level state at $t$ as $s_t := (x, y_{<t})$. Then the following performance-difference identity holds:

**Lemma 1** (Performance difference lemma)**.** *For any two policies $\pi$ and $\mu$, we have*

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}(\rho_t - 1)\Gamma_{t+1}\right], \quad \textit{where } \Gamma_{t+1} := \prod_{j=t+1}^{T}\frac{\pi(y_j|s_j)}{\mu(y_j|s_j)} = \prod_{j=t+1}^{T}\rho_j. \tag{1}$$

The proof is deferred to Appendix B.1. Lemma 1 shows that the policy improvement can be decomposed into token-level policy deviations $\rho_t - 1$, with each deviation weighted by a future likelihood-ratio correction, which we call the *full forward trace*, $\Gamma_{t+1} = \prod_{j=t+1}^{T}\frac{\pi(y_j|s_j)}{\mu(y_j|s_j)}$. Conditional on the prefix up to token $t$, this factor is the likelihood ratio of the remaining tokens $y_{t+1:T}$ under the candidate policy $\pi$ relative to the behavior policy $\mu$.

**Trust region for RLVR.** Though the objective in (1) is exact, directly optimizing it is unstable in long-horizon language generation. The full forward trace $\Gamma_{t+1}(\theta)$ compounds token-level policy mismatch over the future tokens. Thus, a few high-ratio trajectories can dominate the update, leading to high-variance and unstable gradients. To avoid this instability, we typically optimize a *local, $\Gamma_{t+1}$-free* surrogate objective under a trust-region constraint.

**Proposition 1** (Policy improvement)**.** *Suppose $|R(y)| \leqslant \xi$ for all $y$ and $|\rho_t - 1| \leqslant \epsilon$ for all $t$, for some $\epsilon > 0$. Define the population local surrogate as $\mathcal{L}_\mu(\pi) := \mathbb{E}_{y\sim\mu}\big[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\big]$. Then,*

*for any fixed $\pi$,*

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) \geqslant \mathcal{L}_\mu(\pi) - 2\xi T(T-1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2 \tag{2}$$

*holds deterministically. Moreover, for $G$ independent responses sampled from $\mu$, with probability at least $1-\alpha$,*

$$\mathcal{L}_\mu(\pi) \geqslant \widehat{\mathcal{L}}_\mu(\pi) - \xi T\epsilon\sqrt{\frac{2\log(1/\alpha)}{G}},$$

*where the empirical surrogate is defined as $\widehat{\mathcal{L}}_\mu(\pi) := \frac{1}{G}\sum_{i=1}^{G} R(y^{(i)}) \sum_{t=1}^{T}\big(\rho_t^{(i)} - 1\big)$. Combining these inequalities gives the empirical policy-improvement bound with the same probability, where $D_{\mathrm{TV}}^{\max}(\mu,\pi) := \sup_s D_{\mathrm{TV}}(\mu(\cdot|s), \pi(\cdot|s))$.*

The proof is deferred to Appendix B.2. Proposition 1 gives a lower bound on the one-step policy improvement in terms of the empirical local surrogate, a total-variation penalty, and a variance term. The bound implies that increasing $\widehat{\mathcal{L}}_\mu(\pi)$ leads to policy improvement only when the target policy remains sufficiently close to the rollout policy $\mu$, i.e., when $D_{\mathrm{TV}}$ is small. This provides a direct justification for trust-region mechanisms in conventional RLVR: they control the error incurred by replacing the exact policy gradient objective in (1) with the local surrogate. We therefore optimize:

$$\max_{\pi}\ \widehat{\mathcal{L}}_\mu(\pi) = \frac{1}{G}\sum_{i=1}^{G} R(y^{(i)}) \sum_{t=1}^{T}\big(\rho_t^{(i)} - 1\big), \qquad \text{s.t.} \quad D_{\mathrm{TV}}^{\max}(\mu,\pi) \leqslant \delta, \tag{3}$$

By Pinsker's inequality, the TV constraint can also be enforced via a KL-divergence constraint.

**Proximal Policy Optimization (PPO).** Although the constrained objective in (3) has strong theoretical motivation, solving it requires second-order optimization, which is computationally expensive and difficult to scale. PPO [32] was introduced as a practical alternative that preserves the stabilizing effect of trust-region methods through a clipping operation:

$$\widehat{\mathcal{L}}_\mu^{\mathrm{PPO}}(\pi) := \frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{T} \min\left(\rho_t^{(i)}\widehat{A}_t^{(i)},\ \mathrm{clip}\Big(\rho_t^{(i)}, 1-\epsilon, 1+\epsilon\Big)\widehat{A}_t^{(i)}\right), \tag{4}$$

where $\widehat{A}_t^{(i)}$ denotes the estimated advantage at token position $t$ for the $i$-th response. In LLM fine-tuning, the advantage is typically constructed from the centered group reward, e.g., $\widehat{A}_t^{(i)} = \widehat{A}^{(i)} := R(y^{(i)}) - \frac{1}{G}\sum_{j=1}^{G} R(y^{(j)})$ for all $t$ [33, 18]. The clipping operation can be viewed as a *sample-level approximation* to a trust-region constraint. Indeed, for each token state $s_t$, $D_{\mathrm{TV}}(\mu(\cdot|s_t), \pi(\cdot|s_t)) = \frac{1}{2}\mathbb{E}_{y_t\sim\mu(\cdot|s_t)}[|\rho_t - 1|]$. Thus, clipping $\rho_t$ around 1 approximately limits the local policy shift from $\mu$ to $\pi$, providing a practical first-order substitute for the explicit trust-region constraint.

**Masked Policy Gradient (MPG).** Recently, a growing body of work has emphasized that a key factor for stable RL training of LLMs is the control of training–inference mismatch [1, 27, 38, 49, 28]. In particular, masked policy gradient methods—which exclude a small subset of tokens exhibiting large discrepancies in log-probability between the rollout and training policies—have emerged as a general framework for addressing this issue. The objective of MPG is

$$\widehat{\mathcal{L}}_\mu^{\mathrm{MPG}}(\pi) := \frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{T} M_t^{(i)} \cdot \rho_t^{(i)} \cdot \widehat{A}_t^{(i)}, \tag{5}$$

where $M_t^{(i)} \in \{0,1\}$ is a token-level mask. Several existing methods instantiate this template through different choices of $M_t^{(i)}$. For example, GRPO [33] uses the asymmetric ratio-based mask $M_t^{(i)} := \mathbf{1}\{\epsilon_{\mathrm{low}} < \rho_t^{(i)} < \epsilon_{\mathrm{high}} \text{ or } \widehat{A}_t^{(i)}(\rho_t^{(i)} - 1) \leqslant 0\}$, while DPPO [28] replaces this with a divergence-based token mask, e.g., $M_t^{(i)} := \mathbf{1}\{D_{\mathrm{TV}}(\mu(\cdot|s_t^{(i)}), \pi(\cdot|s_t^{(i)})) \leqslant \delta \text{ or } \widehat{A}_t^{(i)}(\rho_t^{(i)} - 1) \leqslant 0\}$. Thus, both methods induce a token-level trust-region effect by retaining update terms only when the policy shift is controlled or when the update moves in the unclipped direction.

## 3 $N$-Step Forward-Trace Policy Optimization

Lemma 1 shows that the full forward trace $\Gamma_{t+1}$ gives an exact, *unbiased* expression for policy improvement. In long-horizon generation, however, this full product can be *unstable* and induce *high variance*, since even modest token-level policy mismatches may compound across the remaining tokens. At the other extreme, ignoring the forward trace entirely gives the *biased* local surrogate underlying standard RLVR objectives such as PPO/GRPO in (4), where stability is enforced only through ratio clipping or token-level divergence masking. Motivated by this bias–variance trade-off, we introduce an **N-step forward trace** that retains only a finite window of future likelihood ratios, interpolating between the local RLVR surrogate and the exact policy gradient objective in (1).

For $N \in \{1, \dots, T\}$, define the $N$-step forward trace $\Gamma_{t+1}^{(N)} := \prod_{j=t+1}^{\min\{t+N-1,T\}} \rho_j$. This correction keeps only the next $N-1$ future likelihood ratios after token $t$. Using this truncated correction, we define the (population) $N$-step surrogate objective as

$$\mathcal{L}_{\mu}^{(N)}(\pi) := \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\Gamma_{t+1}^{(N)}\right]. \tag{6}$$

When $N = 1$, the product is empty, so $\Gamma_{t+1}^{(1)} = 1$. Therefore, the $N$-step surrogate reduces to the local surrogate, i.e., $\mathcal{L}_{\mu}^{(1)}(\pi) = \mathcal{L}_{\mu}(\pi)$. At the other extreme, when $N = T$, the correction includes all future tokens after token $t$, so $\Gamma_{t+1}^{(T)} = \Gamma_{t+1}$. This recovers the exact policy-improvement objective in (1). While larger $N$ may appear preferable since $N = T$ recovers the exact policy gradient objective, this observation applies only to the *population* objective. In practice, optimization is performed using an empirical $N$-step surrogate computed from $G$ sampled responses, namely $\widehat{\mathcal{L}}_{\mu}^{(N)}(\pi) := \frac{1}{G}\sum_{i=1}^{G} R(y^{(i)})\sum_{t=1}^{T}\big(\rho_t^{(i)} - 1\big)\Gamma_{t+1}^{(N,i)}$. As $N$ increases, the variance of this estimator also grows, so the practical behavior is governed by a trade-off between bias and variance.

### 3.1 Theoretical Analysis

The intuition on the bias–variance trade-off of the $N$-step surrogate is formalized in the following theorem, where the proof is provided in Appendix B.3.

**Theorem 1** ($N$-step policy improvement)**.** *Suppose $|R(y)| \leqslant \xi$ for all $y$ and $|\rho_t - 1| \leqslant \epsilon$ for all $t$, for some $\epsilon > 0$. Then, for any fixed $\pi$,*

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) \geqslant \mathcal{L}_{\mu}^{(N)}(\pi) - 2\xi(T-N)(T-N+1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2$$

*holds deterministically. Moreover, for $G$ independent responses sampled from $\mu$, with probability at least $1-\alpha$,*

$$\mathcal{L}_{\mu}^{(N)}(\pi) \geqslant \widehat{\mathcal{L}}_{\mu}^{(N)}(\pi) - B_N\sqrt{\frac{2\log(1/\alpha)}{G}}.$$

*Combining these inequalities gives the empirical policy-improvement bound with the same probability, where $D_{\mathrm{TV}}^{\max}(\mu,\pi) := \sup_s D_{\mathrm{TV}}(\mu(\cdot|s), \pi(\cdot|s))$ and $B_N := \xi\epsilon\sum_{t=1}^{T}(1+\epsilon)^{\min\{N-1,T-t\}}$.*

**Bias–variance trade-off.** At the population level, the $N$-step surrogate $\mathcal{L}_{\mu}^{(N)}(\pi)$ lower bounds the true policy improvement up to a truncation-bias term. This bias decreases with $N$, since a longer forward trace retains more future likelihood-ratio corrections from the exact policy gradient objective in (1). At the finite-sample level, replacing the population surrogate by its empirical counterpart introduces an additional variance penalty of order $\mathcal{O}\big((1+\epsilon)^{N-1}\big)$. Because larger $N$ involves products of more likelihood ratios, this variance term can grow with $N$. Thus, small $N$ gives a stable but more biased objective, whereas large $N$ gives a less biased but potentially less stable objective. Selecting an appropriate intermediate $N$ is therefore essential for achieving consistent and stable policy improvement in practice.

**Bridging local surrogates and exact policy gradient objective.** Theorem 1 makes this interpolation explicit by identifying the two endpoints of the $N$-step surrogate. When $N = 1$, we have $\Gamma_{t+1}^{(1)} = 1$, so the resulting bound reduces to the local surrogate bound in (2), which underlies standard local

surrogate objectives such as PPO/GRPO in (4). When $N = T$, the correction includes all future tokens after token $t$, so $\Gamma_{t+1}^{(T)} = \Gamma_{t+1}$. In this case, the truncation bias vanishes and the bound recovers the exact policy gradient objective in (1). Thus, the proposed $N$-step formulation provides a continuous bridge between local policy optimization and exact policy improvement.

**Why choosing the right trace horizon matters.** Theorem 1 suggests that neither endpoint should be used by default. The trace horizon $N$ is a central design parameter: it determines how much future correction is incorporated into the objective, and therefore directly affects the balance between bias reduction and optimization stability. Choosing $N$ too small may discard useful future correction and leave substantial truncation bias; choosing $N$ too large may introduce unstable long-range likelihood-ratio products and inflate the finite-sample penalty. Thus, the best practical choice is often an intermediate trace horizon rather than either endpoint $N = 1$ or $N = T$.

We therefore treat $N$ as an important hyperparameter. Larger values of $N$ are preferable when the updated policy remains close to the behavior policy (i.e., small $D_{\mathrm{TV}}^{\max}$) and when token-level likelihood ratios are stable (i.e., small $\epsilon$). Smaller values of $N$ are preferable when the update approaches the trust-region boundary (i.e., large $D_{\mathrm{TV}}^{\max}$) or future-token ratios become unstable (i.e., large $\epsilon$). Selecting $N$ appropriately is therefore essential for obtaining an objective that is both stable and accurate enough to translate into strong empirical performance.

### 3.2 Toy Example: Small Token MDP

To illustrate the bias-variance trade-off of $N$-step surrogate, we consider a token MDP with a small vocabulary $\mathcal{V} = \{a, b, c\}$ and a fixed horizon $T = 7$. A trajectory $(y_1, \ldots, y_T)$ is generated by a behavior policy $\mu$ and evaluated under a target policy $\pi$. The reward is defined as a binary function of the completed sequence, indicating whether a predefined target pattern appears as an ordered subsequence. Further details on the token MDP can be found in Appendix C. The results clearly exhibit the predicted trade-off. As $N$ increases, the population surrogate objective $\mathcal{J}(\mu) + \mathcal{L}_{\mu}^{(N)}(\pi)$ approaches the true objective $\mathcal{J}(\pi)$, confirming that the truncation bias decreases with $N$.

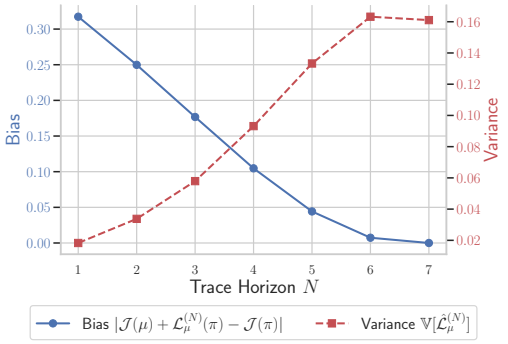



Figure 2: Bias–variance trade-off of the $N$-step surrogate objective on the token MDP example.

However, at the same time, the variance $\mathbb{V}[\hat{\mathcal{L}}_{\mu}^{(N)}(\pi)]$ increases with $N$, reflecting the growing instability introduced by longer products of importance ratios. Overall, this example provides a clean and controlled validation of the theory: small $N$ yields a stable but biased objective, while large $N$ reduces bias at the cost of increased variance.

### 3.3 Practical Implementation

We now introduce **N-Step Forward-Trace Policy Optimization (NFPO)**, a practical RLVR algorithm based on the $N$-step forward trace. As indicated by Theorem 1, a trust region is required for the $N$-step surrogate objective in (6) to guarantee policy improvement. To enable a practical implementation of the trust region, we build upon the masked policy gradient framework and incorporate an $N$-step forward trace. In practice, we apply stop-gradient to the forward trace for numerical stability, which leads to an objective based on $\rho_t$ rather than $(\rho_t - 1)$:

$$\widehat{\mathcal{L}}_{\mu,\mathrm{NFPO}}^{(N)}(\pi) := \frac{1}{G} \sum_{i=1}^{G} \sum_{t=1}^{T} M_t^{(i)} \cdot \widehat{A}^{(i)} \cdot \rho_t^{(i)} \cdot \bar{\Gamma}_{t+1}^{(N,i)},$$

where $\bar{\Gamma}_{t+1}^{(N,i)}$ denotes the clipped $N$-step forward trace defined below. To satisfy the bounded-ratio condition in Theorem 1 while ensuring numerical stability, we clip the token-level likelihood ratio as

$\rho_{t,\beta}^{(i)} := \mathrm{clip}\left(\rho_t^{(i)}, 1/\beta, \beta\right)$. Based on this, we define the clipped $N$-step forward trace as

$$\bar{\Gamma}_{t+1}^{(N,i)} := \mathrm{clip}\left(\prod_{j=t+1}^{m_t(N)} \rho_{j,\beta}^{(i)}, 1-\epsilon_{\mathrm{low}}, 1+\epsilon_{\mathrm{high}}\right).$$

Since the lower bound in Theorem 1 is expressed in terms of a total variation constraint, we adopt the token mask based on total variation from DPPO [28], where $M_t^{(i)} := \mathbf{1}\{D_{\mathrm{TV}}\big(\mu(\cdot|s_t^{(i)}), \pi(\cdot|s_t^{(i)})\big) \leqslant \delta \text{ or } \widehat{A}^{(i)}\big(\rho_t^{(i)} - 1\big) \leqslant 0\}$. Following DPPO, we directly use the log probabilities from the rollout engine as $\mu$ without recomputing them in the training engine. Thus, $\rho_t^{(i)}$ also corrects the mismatch between the rollout and training engines. The $N$-step forward trace is then computed by multiplying the clipped ratios $\rho_{j,\beta}^{(i)}$ of the subsequent $N-1$ tokens, which are already computed for the surrogate objective. Hence, NFPO requires no additional forward or backward passes, and the additional cost is negligible (Appendix E.4).

## 4 Main Experiments on LLM Reasoning Benchmarks

In this section, we empirically evaluate NFPO on LLM reasoning tasks.

**Models and baselines.** We employ Qwen3-1.7B-Base [44] and Qwen3-8B-Base as base models. We consider three RLVR algorithms: GRPO, the standard RLVR algorithm that utilizes group-relative advantage estimation and PPO-style clip; GSPO [50], which uses sequence-level importance ratios and clipping; and DPPO, which applies divergence-aware token mask to implement trust-region.

**Training and evaluation.** For mathematical reasoning, we use the level 3-5 problems of MATH [12] dataset with 8k maximum response length for Qwen3-1.7B-Base, and DAPO-Math-17k dataset [45] with 16k response length for Qwen3-8B-Base. We sample 128 prompts per step, and generate 8 responses per prompt. After generating the 1024 responses, the policy is updated using mini-batches of size 256. The performance is measured by seven mathematical reasoning benchmarks: AIME24/25/26 [22], AMC23 [21], MATH 500 [12, 16], Minerva [14], and OlympiadBench (Oly.) [10]. We report the pass@1 metric averaged over 32 responses per prompt, with sampling temperature $0.7$. Beyond mathematical reasoning, we further train on four additional domains: logical reasoning (Countdown [25]), code generation (PRIME Code [5]), multi-turn agentic tasks (ALFWorld [35]), and multi-turn puzzles (Sudoku [19]). We use Qwen3-1.7B-Base for Countdown, PRIME Code, and Sudoku, and Qwen2.5-1.5B-Instruct for ALFWorld, with the same configuration of $N = 4$ and a forward-trace clipping range of $[0.8, 1.2]$ in all four domains. Since DPPO achieves the strongest performance among the baselines on mathematical reasoning, we use DPPO as the baseline in these domains. More details are in Appendix C and Appendix E.2.

**Performance evaluation.** Table 1 presents the evaluation results on comprehensive math reasoning benchmarks. NFPO consistently outperforms or matches GRPO, GSPO, and DPPO in almost every benchmark. In particular, Qwen3-1.7B-Base model trained by NFPO achieves $30.7\%$ accuracy on average, which is $16.3\%$ improvement compared to GRPO, and $10.0\%$ compared to DPPO. The same trend persists in the scaling experiment with Qwen3-8B-Base model, where NFPO continues to outperform the baselines across most evaluation benchmarks. Table 2 presents the evaluation results on the four additional domains, where NFPO outperforms DPPO on every metric. In particular, NFPO improves pass@1 by $11.9\%$, $11.4\%$, and $12.4\%$ compared to DPPO on Countdown, PRIME Code, and ALFWorld, respectively, and more than doubles the episode return on Sudoku. These results suggest that the advantage of forward trace correction remains consistent across model scales and diverse reasoning benchmarks, and is not specific to mathematical reasoning or domain-specific tuning. Detailed training metrics and validation curves are presented in Appendix E.

## 5 Learning-Dynamics Analysis of Forward Trace Correction

We further explore the effect of the forward trace on RL training. This empirical analysis augments the theoretical findings in Section 3.1 by revealing how forward trace changes the training dynamics.

**Observation 1: NFPO corrects tokens more actively, especially reasoning tokens.** Our token analysis shows that the forward trace enables more active bias correction, and the effect is more pronounced for *reasoning tokens* such as "suppose", "since", and "wait".

Table 1: Performance comparison on mathematical reasoning benchmarks. We report pass@1 averaged over 32 responses per prompt.

| | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| Qwen3-1.7B-Base | 0.034 | 0.015 | 0.021 | 0.261 | 0.485 | 0.134 | 0.184 | 0.162 |
| GRPO | 0.094 | 0.050 | 0.047 | 0.425 | 0.690 | 0.225 | 0.315 | 0.264 |
| GSPO | 0.098 | 0.055 | 0.053 | 0.416 | 0.692 | 0.224 | 0.324 | 0.266 |
| DPPO | 0.106 | 0.066 | 0.056 | 0.424 | 0.715 | 0.246 | 0.341 | 0.279 |
| NFPO (ours) | **0.133** | **0.074** | **0.081** | **0.496** | **0.745** | **0.262** | **0.360** | **0.307** |
| Qwen3-8B-Base | 0.094 | 0.074 | 0.056 | 0.423 | 0.623 | 0.219 | 0.292 | 0.254 |
| GRPO | 0.274 | 0.229 | 0.230 | 0.741 | 0.872 | 0.351 | 0.527 | 0.461 |
| GSPO | 0.291 | 0.239 | 0.258 | 0.715 | 0.870 | 0.371 | 0.531 | 0.468 |
| DPPO | 0.347 | 0.256 | 0.265 | 0.791 | 0.886 | **0.399** | 0.565 | 0.501 |
| NFPO (ours) | **0.364** | **0.284** | **0.286** | **0.820** | **0.892** | 0.390 | **0.571** | **0.515** |

Table 2: Results on four additional domains. We report pass@1 for Countdown, PRIME Code, and ALFWorld, and the episode return for Sudoku.

| | Countdown<br>Logical reasoning | PRIME Code<br>Code generation | ALFWorld<br>Multi-turn agentic | Sudoku<br>Multi-turn puzzle |
|---|---|---|---|---|
| DPPO | 0.468 | 0.236 | 0.765 | 0.318 |
| NFPO (ours) | **0.524** | **0.263** | **0.860** | **0.656** |

The likelihood ratio $\rho_t$ corresponds to the token-level weight of standard PPO surrogate objectives (e.g. GRPO), whereas NFPO augments each weight with the forward trace, yielding *trace-corrected likelihood ratio* $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$. During the RL training, tokens whose token-level weight deviates more from 1 receive more active correction signals. To examine how forward trace changes the dynamics of token-level correction, we measure correction strength for the two objectives: $\mathbb{E}_t[|\rho_t - 1|]$ represents the correction strength of the PPO surrogate objective, whereas $\mathbb{E}_t[|\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)} - 1|]$ quantifies the correction strength of NFPO. We compute correction strength for approximately $3.4$M tokens collected during training, and categorize the tokens semantically to compare the correction strength across categories (see Appendix D for details of the procedure).

Table 3: Correction strength of likelihood ratio and trace-corrected likelihood ratio, measured for each token category.

| Category | Correction strength | |
|---|---|---|
| | $\rho_t$ | $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$ |
| All tokens | 0.0278 | 0.0722 |
| Reasoning tokens | 0.0531 | 0.1078 |

As shown in Table 3, the correction strength of the trace-corrected likelihood ratio is substantially higher than that of the uncorrected likelihood ratio. This implies that NFPO more actively applies correction signals to reduce the bias from policy deviation. In contrast, $\rho_t$ alone provides weak correction signals. Moreover, the correction strength for reasoning tokens is much higher than the average for both objectives, showing that reasoning tokens are critical for bias correction. This finding aligns with prior work showing that reasoning performance in RLVR is driven by these reasoning tokens that act as forks in the reasoning trajectory [41]. In summary, the forward trace $\bar{\Gamma}_{t+1}^{(N,i)}$ enables more active bias correction, and such correction signals are even intensified for reasoning tokens. This explains why NFPO is more effective in learning complex reasoning tasks, under policy deviation.

**Observation 2: Forward trace acts as low-pass filter.** The key effect of the $N$-step forward trace is to smooth noisy token-wise likelihood-ratio corrections into coherent segment-level update signals. As shown in the left panel of Figure 3, the token-level likelihood ratio $\rho_t$ can oscillate rapidly around the neutral value of 1 across neighboring token positions. Such high-frequency switching can harm optimization because it breaks local coherence and distorts the policy gradient updates of neighboring tokens, ultimately destabilizing training [11]. In contrast, NFPO uses the trace-corrected ratio $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$, where the forward trace aggregates the likelihood ratios of the subsequent $N-1$ tokens. This aggregation couples the correction signals of nearby tokens, since adjacent positions

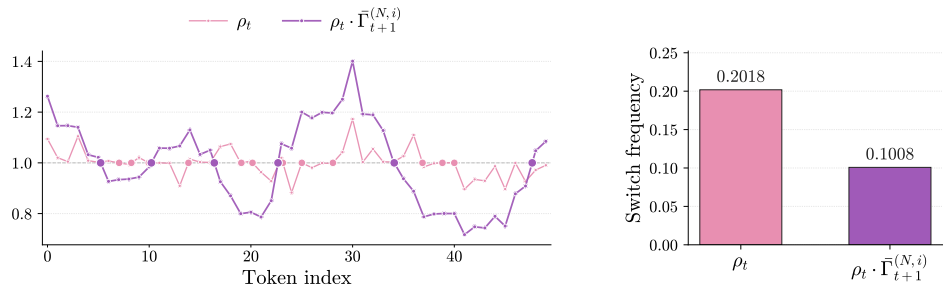



Figure 3: (Left) Comparison of token-level ($\rho_t$) and trace-corrected ($\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$) likelihood ratios along a 50-token rollout segment. (Right) Batch-averaged switch frequency, halved by the forward trace.

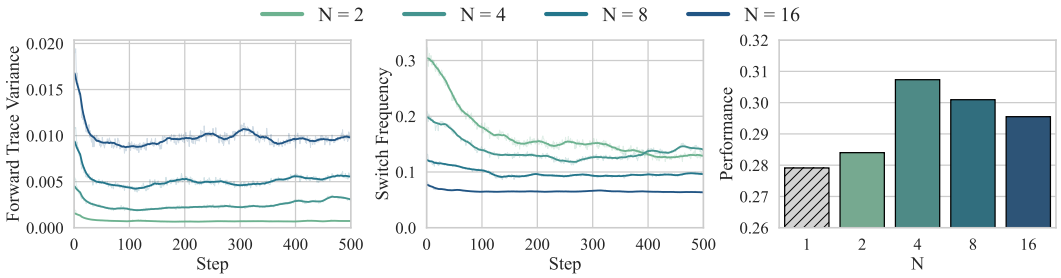



Figure 4: Ablation results across $N$, where the $N = 1$ corresponds to DPPO. (Left) Variance of forward trace. (Middle) Switch frequency of forward trace. (Right) Performance on the benchmarks.

share most of their $N$-step forward trace. As a result, isolated token-level fluctuations are suppressed, while persistent signals over local segments are preserved.

To quantify this smoothing effect, we measure the switch frequency, defined as the fraction of adjacent token pairs whose ratio signals lie on opposite sides of the neutral value 1. A high switch frequency indicates rapid alternation between up-weighting and down-weighting neighboring tokens, while a low switch frequency reflects more coherent segment-level correction. As shown in the right panel of Figure 3, the forward trace roughly halves the batch-averaged switch frequency, reducing it from $20.18\%$ for $\rho_t$ to $10.08\%$ for $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$. This confirms that the forward trace suppresses high-frequency token-level oscillations and acts as a low-pass filter over likelihood-ratio corrections.

**Observation 3: An intermediate $N$ achieves the best performance.** We analyze the effect of $N$, which determines the accumulation length of the forward trace. The results in Figure 4 reveal a clear pattern. In terms of performance, $N = 4$ emerges as a *sweet spot*, while all configurations with $N > 1$ consistently outperform the $N = 1$ baseline.

The existence of such a sweet spot becomes evident when examining the training dynamics. As $N$ grows, the variance of the forward trace increases. While a moderate level of variance reflects active correction and is therefore beneficial, excessive variance can introduce noise into the gradient estimates and degrade training stability. At the same time, increasing $N$ enhances semantic low-pass filtering, which is indicated by the reduced switch frequency of $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N)}$. These two effects evolve in opposing directions as $N$ varies, giving rise to an intermediate regime where the variance in forward trace and the strength of the low-pass filtering are optimally balanced. This interplay naturally leads to the emergence of a performance sweet spot.

## 6 Ablation Study

We present a comprehensive ablation study on the hyperparameters and design choices of NFPO.

**Ablation 1. Forward trace clip range.** We investigate the impact of the clip range of forward trace. Using the clip range of GRPO as a reference, we start with the $[0.8, 1.2]$ range and compare with the clip-higher (increasing $\epsilon_{\text{high}}$ to $1.4$) and clip-lower (decreasing $\epsilon_{\text{low}}$ to $0.6$). The left plot in Figure 5

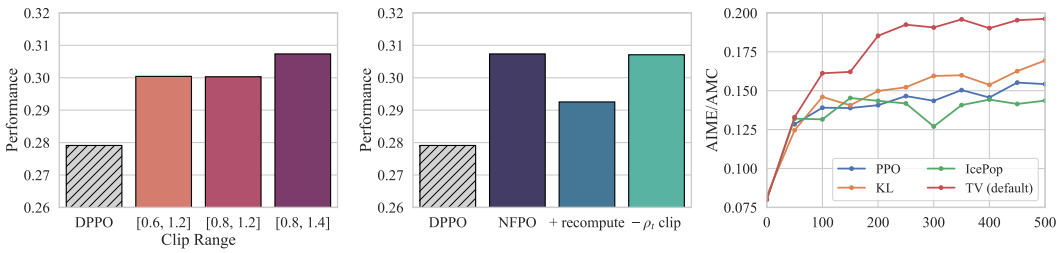



Figure 5: Ablation experiments on NFPO: (Left) Different clip ranges, (Middle) Alternative design choices, and (Right) Different token masks.

shows that all considered clipping ranges outperform the DPPO baseline, demonstrating that the performance improvement of NFPO is robust to the clip range.

**Ablation 2. Log probability recompute.** Training–inference mismatch introduces a design choice regarding whether to recompute the rollout policy's log probabilities within the training backend. Since forward trace computation also relies on the rollout log probabilities, we evaluate the effect of recomputing them during training. The middle plot in Figure 5 shows that recompute leads to a slight performance degradation. Considering the extra computational overhead required, we implement NFPO without log probability recompute by default.

**Ablation 3. Likelihood ratio clip.** In our implementation of NFPO, we clip the likelihood ratio $\rho_t$ for numerical stability. This clipping serves as a minimal safeguard to prevent a small number of outliers from disrupting the training dynamics; accordingly, the clipping threshold is fixed at $\beta = 3.0$ across all experiments without tuning. The middle plot in Figure 5 shows that removing this clipping still yields a similar performance. This behavior is likely due to the clipping applied to the forward trace itself, which prevents variance from exploding, and suggests that the effectiveness of the forward trace is not particularly sensitive to clipping applied to $\rho_t$.

**Ablation 4. Token mask.** NFPO is built upon the masked policy gradient framework, where the trust region is enforced via token-level masking. In our main experiments, we adopt the token mask from DPPO, which implements a trust region defined by total variation. However, NFPO is not restricted to this specific choice. The right plot in Figure 5 presents the performance of NFPO with various token masks. The total variation-based mask [28] outperforms the KL divergence based mask [28], the PPO clipping used in GRPO, and the token mask employed in IcePop [38]. Accordingly, we adopt the TV–based token mask as the default choice.

## 7 Conclusion

This work introduces the concept of the $N$-step forward trace, which bridges the widely used local surrogate objectives in RLVR and the exact policy gradient objective. We show that the local surrogate discards the forward trace in the exact policy improvement identity, and that the $N$-step forward trace reduces this truncation bias at the cost of higher variance. Thus, forward trace correction provides a principled mechanism for interpolating between local surrogate losses and exact policy gradients, where the trace horizon $N$ controls the bias–variance trade-off. Building on this analysis, we propose NFPO, which implements the $N$-step forward trace within a masked policy gradient framework.

Through both theoretical analysis and large-scale LLM experiments, we demonstrate that an appropriate choice of the trace horizon $N$ can consistently improve performance over conventional PPO/GRPO-style objectives. NFPO consistently outperforms GRPO, GSPO, and DPPO on seven mathematical reasoning benchmarks across two model scales, and also outperforms DPPO on logical reasoning, code generation, multi-turn agentic tasks, and multi-turn puzzles. Furthermore, our mechanistic analyses provide intuitive insight into how forward trace correction improves learning dynamics in practice. We find that the forward trace concentrates updates on reasoning tokens and acts as a low-pass filter on noisy token-level likelihood ratios. Experiments varying $N$ further support the predicted bias–variance trade-off, complementing and reinforcing our theoretical findings.

## Acknowledgments

This work was supported by Upstage. DY, JL, BK and MO were supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant, both funded by the Korean government (MSIT) (No. RS-2022-NR071853, RS-2023- 00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008, RS-2025-25420849). JL was additionally supported by the 2025 Google PhD Fellowship, with support from Google.org.

## References

[1] 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.

[2] Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. Pass@ k training for adaptively balancing exploration and exploitation of large reasoning models. *arXiv preprint arXiv:2508.10751*, 2025.

[3] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30377–30385, 2026.

[4] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=inccdtfx8x`.

[5] Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

[6] 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*, 2025.

[7] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[8] 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.

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, 2024.

[11] Shuo He, Lang Feng, Xin Cheng, Lei Feng, and Bo An. Online causal kalman filtering for stable and effective policy optimization. *arXiv preprint arXiv:2602.10609*, 2026.

[12] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL `https://openreview.net/forum?id=7Bywt2mQsCe`.

[13] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

[14] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[15] Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. Beyond Pass@1: Self-play with variational problem synthesis sustains RLVR. *arXiv preprint arXiv:2508.14029*, 2025.

[16] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

[17] ZhiHang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. CPPO: Accelerating the training of group relative policy optimization-based reasoning models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=SVHerutWxp.

[18] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[19] Zichen Liu, Anya Sims, Keyu Duan, Changyu Chen, Simon Yu, Xiangxin Zhou, Haotian Xu, Shaopan Xiong, Bo Liu, Chenmien Tan, Chuen Yang Beh, Weixun Wang, Hao Zhu, Weiyan Shi, Diyi Yang, Michael Shieh, Yee Whye Teh, Wee Sun Lee, and Min Lin. GEM: A gym for agentic LLMs. *arXiv preprint arXiv:2510.01051*, 2025.

[20] Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. Fipo: Eliciting deep reasoning with future-kl influenced policy optimization. *arXiv preprint arXiv:2603.19835*, 2026.

[21] MAA. American mathematics competitions (AMC). https://maa.org/, 2023. MAA. American mathematics competitions–AMC.

[22] MAA. American invitational mathematics examination (AIME). https://maa.org/, 2026. MAA. American invitational mathematics examination–AIME.

[23] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024. Blog.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[25] Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025.

[26] Aleksei Petrenko, Ben Lipkin, Kevin Chen, Erik Wijmans, Marco Cusumano-Towner, Raja Giryes, and Philipp Krähenbühl. Entropy-preserving reinforcement learning. *arXiv preprint arXiv:2603.11682*, 2026.

[27] Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.

[28] Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

[29] Yun Qu, Qi Wang, Yixiu Mao, Heming Zou, Yuhang Jiang, Yingyue Li, Wutong Xu, Lizhou Cai, Weijie Liu, Clive Bai, Kai Yang, Yangkun Chen, Saiyong Yang, and Xiangyang Ji. Listwise policy optimization: Group-based RLVR as target-projection on the LLM response simplex. *arXiv preprint arXiv:2605.06139*, 2026.

[30] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[31] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

[33] 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.

[34] 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*, pages 1279–1297, 2025.

[35] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

[36] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

[37] Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. Maximum likelihood reinforcement learning. *arXiv preprint arXiv:2602.02710*, 2026.

[38] Ling Team, Anqi Shen, Baihui Li, Bin Hu, Bin Jing, Cai Chen, Chao Huang, Chao Zhang, Chaokun Yang, Cheng Lin, et al. Every step evolves: Scaling reinforcement learning for trillion-scale thinking model. *arXiv preprint arXiv:2510.18855*, 2025.

[39] Christian Walder and Deep Karkhanis. Pass@K policy optimization: Solving harder reinforcement learning problems. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[40] Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

[41] Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 39, 2025.

[42] Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation for entropy-safe reasoning. *arXiv preprint arXiv:2509.22611*, 2025.

[43] Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, et al. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*, 2025.

[44] 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*, 2025.

[45] 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. *arXiv preprint arXiv:2503.14476*, 2025.

[46] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[47] Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*, 2025.

[48] Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, et al. Geometric-mean policy optimization. *arXiv preprint arXiv:2507.20673*, 2025.

[49] Chujie Zheng, Kai Dang, Bowen Yu, Mingze Li, Huiqiang Jiang, Junrong Lin, Yuqiong Liu, Hao Lin, Chencan Wu, Feng Hu, et al. Stabilizing reinforcement learning with llms: Formulation and practices. *arXiv preprint arXiv:2512.01374*, 2025.

[50] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[51] Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=ftVlLG9cks`.

# A Related Works

**Reinforcement learning with verifiable rewards (RLVR).** Reinforcement learning has become a central paradigm for post-training large language models [24, 30, 23, 9]. In particular, for reasoning tasks such as mathematics and coding, reinforcement learning with verifiable rewards (RLVR)—where rewards are assigned based on objective, programmatically verifiable outcomes—has demonstrated notable success. Building on the theoretical foundations of Trust Region Policy Optimization (TRPO) [31] and Proximal Policy Optimization (PPO) [32], a range of methods tailored to large-scale language model training have been developed.

Group Relative Policy Optimization (GRPO) [33, 9] eliminates the need for a learned critic and instead employs group-relative advantage estimation. This approach has been shown to achieve strong empirical performance while reducing cost, and has since emerged as a standard framework in RLVR training of reasoning language models. Subsequent works have further improved this framework along multiple directions, including more effective advantage computation [45, 17, 42, 3], sampling strategies [45, 47, 43, 51], entropy regularization [6, 26], enhanced training stability [50, 1, 38, 49, 27, 48, 8, 28], and improved learning objective [18, 4, 40, 2, 37].

**Surrogate objectives for policy gradient.** When the behavior policy used for data collection is close to the training policy, the policy gradient objective can be locally approximated, yielding a surrogate objective that enables off-policy optimization without requiring data to be regenerated under the evolving policy. Conservative Policy Iteration (CPI) [13] derives policy improvement guarantees using such a surrogate objective under a policy-mixture update rule, while TRPO [31] enforces a trust-region constraint to ensure the validity of the local approximation during policy updates. However, the update rule of CPI is difficult to implement for complex policy classes, and although TRPO is more practical, it still requires second-order optimization.

PPO [32] implements the trust-region concept of TRPO via simple clipping operations, making it applicable to complex policy classes. As a result, most algorithms used in RLVR training of large language models, including GRPO [33, 9], are developed as variants of PPO. To address domain-specific challenges in LLM training, several clipping strategies and token-level masking techniques have been proposed [45, 50, 8, 38, 48, 28]. Nevertheless, the majority of these methods are fundamentally built upon the surrogate objective derived from CPI and TRPO, i.e., a locally approximated policy gradient objective.

In this work, we revisit this local approximation and show that a surrogate objective corrected by an $N$-step forward trace bridges the fully local surrogate and the exact policy gradient. Our contribution therefore does not lie in refining the implementation of PPO-style objectives, but in modifying the objective itself to achieve improved performance.

**Connection with FIPO [20].** Very recently, Ma et al. [20] proposed Future-KL Influenced Policy Optimization (FIPO), a future-dependent token reweighting method that accumulates *discounted* log-probability shifts over the remaining tokens to modulate token-level policy updates. This also induces a future-dependent multiplicative weight on token-level updates, making FIPO superficially related to our forward trace formulation. The key difference is conceptual and theoretical. FIPO is motivated as a *heuristic* dense-credit-assignment rule based on future-KL signals, while our method is derived from the *exact policy-improvement identity*, where the forward trace appears as the principled likelihood-ratio correction term. Consequently, our objective is not designed as an ad hoc advantage reweighting scheme; rather, it corrects the structural bias of local RLVR objectives and exposes a formal bias–variance trade-off governed by the trace horizon $N$.

Moreover, when the discount factor in FIPO is set to one, its future-dependent reweighting can be interpreted as the full-horizon instance of our framework, corresponding to the forward trace with $N = T$. This highlights the generality of our formulation: it explains existing future-dependent reweighting structures while also providing a principled truncated family with an explicit bias–variance characterization.

For completeness, we also compare our method empirically with FIPO. The two methods differ not only in how future-dependent signals are incorporated, but also in implementation details such as the trust-region mechanism. As shown in Figure A.1, our method consistently outperforms FIPO by a substantial margin. Rather than attributing this gap to a single isolated factor, we view it as the

result of a theoretically grounded design that better realizes the benefits of forward trace correction. In particular, the masked policy-gradient framework with a total-variation-based token mask provides a reliable way to enforce the trust-region condition in Theorem 1.

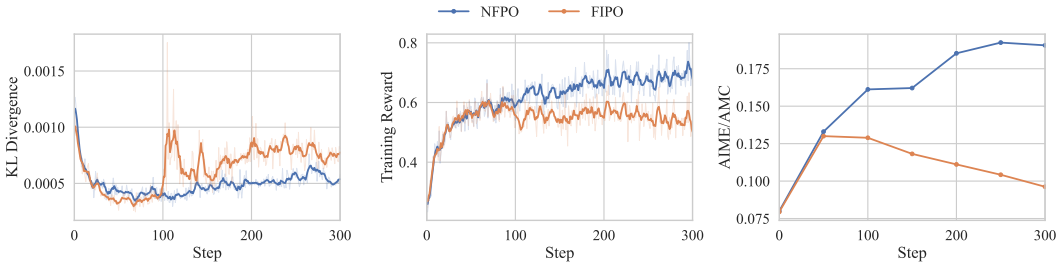



Figure A.1: (Left) KL divergence between training and rollout policy, (middle) training reward, and (right) average evaluation performance on AIME24/25/26 and AMC23 for NFPO and FIPO. The base model is Qwen3-1.7B-Base model, and the training setting follows that described in Section 4.

# B Detailed Proofs

## B.1 Proof of Performance Difference Lemma

*Proof of Lemma 1.* By definition, the performance difference can be written as

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \sum_y \big(\pi(y) - \mu(y)\big) R(y). \tag{B.1}$$

Using the autoregressive factorization, we have $\pi(y) = \prod_{t=1}^{T} \pi(y_t|s_t)$ and $\mu(y) = \prod_{t=1}^{T} \mu(y_t|s_t)$. Then, by the telescoping identity, we get

$$\pi(y) - \mu(y) = \sum_{t=1}^{T} \left( \prod_{k=1}^{t-1} \mu(y_k|s_k) \right) \big(\pi(y_t|s_t) - \mu(y_t|s_t)\big) \left( \prod_{j=t+1}^{T} \pi(y_j|s_j) \right).$$

Substituting this identity into (B.1) yields

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \sum_y R(y) \sum_{t=1}^{T} \left( \prod_{k=1}^{t-1} \mu(y_k|s_k) \right) \big(\pi(y_t|s_t) - \mu(y_t|s_t)\big) \left( \prod_{j=t+1}^{T} \pi(y_j|s_j) \right).$$

Now multiply and divide by $\mu(y_t|s_t) \prod_{j=t+1}^{T} \mu(y_j|s_j)$, and note that

$$\mu(y) = \left( \prod_{k=1}^{t-1} \mu(y_k|s_k) \right) \mu(y_t|s_t) \prod_{j=t+1}^{T} \mu(y_j|s_j).$$

Hence,

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \sum_y \mu(y) R(y) \sum_{t=1}^{T} \left( \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)} - 1 \right) \prod_{j=t+1}^{T} \frac{\pi(y_j|s_j)}{\mu(y_j|s_j)}.$$

Using the definitions $\rho_t = \frac{\pi(y_t|s_t)}{\mu(y_t|s_t)}$ and $\Gamma_{t+1} = \prod_{j=t+1}^{T} \rho_j$, we obtain

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \mathbb{E}_{y\sim\mu} \left[ R(y) \sum_{t=1}^{T} (\rho_t - 1) \Gamma_{t+1} \right],$$

which proves the claim. □

### B.2 Proof of Policy Improvement

*Proof of Proposition 1.* The first inequality follows directly from Theorem 3.2 of [28]; we include the proof here for completeness. Fix a policy $\pi$, and write $\delta := D_{\mathrm{TV}}^{\max}(\mu, \pi) = \sup_s D_{\mathrm{TV}}(\mu(\cdot|s), \pi(\cdot|s))$. By the performance difference lemma (Lemma 1), we have

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\Gamma_{t+1}\right],$$

where $\Gamma_{t+1} := \prod_{j=t+1}^{T}\rho_j$. Therefore, we get

$$\begin{aligned}\mathcal{J}(\pi) - \mathcal{J}(\mu) &= \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\right] + \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\big(\Gamma_{t+1} - 1\big)\right] \\ &= \mathcal{L}_\mu(\pi) + \mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\big(\Gamma_{t+1} - 1\big)\right]. \end{aligned} \tag{B.2}$$

Using $|R(y)| \leqslant \xi$, it remains to bound the second term from below:

$$\mathbb{E}_{y\sim\mu}\left[R(y)\sum_{t=1}^{T}\big(\rho_t - 1\big)\big(\Gamma_{t+1} - 1\big)\right] \geqslant -\xi\sum_{t=1}^{T}\mathbb{E}_{y\sim\mu}\left[|\rho_t - 1|\,|\Gamma_{t+1} - 1|\right]. \tag{B.3}$$

We first record two elementary bounds. For any state $s$,

$$\mathbb{E}_{a\sim\mu(\cdot|s)}\left[\left|\frac{\pi(a|s)}{\mu(a|s)} - 1\right|\right] = \sum_a |\pi(a|s) - \mu(a|s)| = 2D_{\mathrm{TV}}(\mu(\cdot|s), \pi(\cdot|s)) \leqslant 2\delta. \tag{B.4}$$

Next, conditional on $s_{t+1} = (x, y_{\leqslant t})$, the forward trace $\Gamma_{t+1}(\theta)$ is the likelihood ratio of the future-token sequence $y_{t+1:T}$ under $\pi$ relative to $\mu$. Hence,

$$\mathbb{E}_\mu\left[|\Gamma_{t+1} - 1| \mid s_{t+1}\right] = 2D_{\mathrm{TV}}\big(P_\pi^{t+1:T}(\cdot|s_{t+1}), P_\mu^{t+1:T}(\cdot|s_{t+1})\big).$$

By a standard telescoping/coupling bound for sequential distributions, we get

$$D_{\mathrm{TV}}\big(P_\pi^{t+1:T}(\cdot|s_{t+1}), P_\mu^{t+1:T}(\cdot|s_{t+1})\big) \leqslant \sum_{j=t+1}^{T} D_{\mathrm{TV}}^{\max}(\mu, \pi) = (T-t)\delta.$$

Therefore, we obtain

$$\mathbb{E}_\mu\left[|\Gamma_{t+1} - 1| \mid s_{t+1}\right] \leqslant 2(T-t)\delta. \tag{B.5}$$

Combining the two bounds, for each $t \leqslant T-1$, we have

$$\begin{aligned}\mathbb{E}_\mu\left[|\rho_t - 1|\,|\Gamma_{t+1} - 1|\right] &= \mathbb{E}_\mu\left[|\rho_t - 1|\,\mathbb{E}_\mu\left[|\Gamma_{t+1} - 1| \mid s_{t+1}\right]\right] \\ &\leqslant 2(T-t)\delta\,\mathbb{E}_\mu\left[|\rho_t - 1|\right] && \text{(Eqn. (B.5))} \\ &\leqslant 4(T-t)\delta^2. && \text{(Eqn. (B.4))}\end{aligned}$$

For $t = T$, let $\Gamma_{T+1} = 1$, so the corresponding term is zero. Thus,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}_\mu\left[|\rho_t - 1|\,|\Gamma_{t+1} - 1|\right] &\leqslant 4\delta^2\sum_{t=1}^{T-1}(T-t) \\ &= 4\delta^2\cdot\frac{T(T-1)}{2} \\ &= 2T(T-1)\delta^2.\end{aligned}$$

Combining this bound with (B.2) and (B.3) gives

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) \geqslant \mathcal{L}_\mu(\pi) - 2\xi T(T-1)\Big(D_{\mathrm{TV}}^{\max}(\mu, \pi)\Big)^2, \tag{B.6}$$

which proves the first lower bound.

It remains to relate the population surrogate to its empirical counterpart. Define the single-sample surrogate estimator

$$Z_\pi(y) := R(y) \sum_{t=1}^{T} \big(\rho_t - 1\big).$$

Then $\mathcal{L}_\mu(\pi_\theta) = \mathbb{E}_\mu[Z_\pi(y)]$ and $\widehat{\mathcal{L}}_\mu(\pi_\theta) = \frac{1}{G}\sum_{i=1}^{G} Z_\pi(y^{(i)})$. By the bounded-reward assumption and the likelihood-ratio deviation bound, we get

$$|Z_\pi(y)| = \left| R(y) \sum_{t=1}^{T} \big(\rho_t - 1\big) \right| \leqslant |R(y)| \sum_{t=1}^{T} |\rho_t - 1| \leqslant \xi T \epsilon.$$

Since $\{y^{(i)}\}_{i=1}^{G}$ are i.i.d. samples from $\mu(\cdot|x)$, the random variables $\{Z_\pi(y^{(i)})\}_{i=1}^{G}$ are also i.i.d. and bounded. By Hoeffding's inequality, with probability at least $1-\alpha$, we have

$$\mathbb{E}_\mu[Z_\pi(y)] \geqslant \frac{1}{G}\sum_{i=1}^{G} Z_\pi(y^{(i)}) - \xi T \epsilon \sqrt{\frac{2\log(1/\alpha)}{G}}.$$

Equivalently,

$$\mathcal{L}_\mu(\pi) \geqslant \widehat{\mathcal{L}}_\mu(\pi) - \xi T \epsilon \sqrt{\frac{2\log(1/\alpha)}{G}}.$$

Combining this concentration bound with the deterministic lower bound in (B.6), we obtain

$$\begin{aligned}
\mathcal{J}(\pi) - \mathcal{J}(\mu) &\geqslant \mathcal{L}_\mu(\pi) - 2\xi T(T-1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2 \\
&\geqslant \widehat{\mathcal{L}}_\mu(\pi) - 2\xi T(T-1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2 - \xi T \epsilon \sqrt{\frac{2\log(1/\alpha)}{G}}.
\end{aligned}$$

This completes the proof of Proposition 1. □

### B.3 Proof of Theorem 1

*Proof of Theorem 1.* Fix a target policy $\pi$ and a forward trace horizon $N \in \{1, \dots, T\}$. We prove the result in three steps.

**Step 1: Population truncation error.** Recall that, by Lemma 1, we have

$$\mathcal{J}(\pi) - \mathcal{J}(\mu) = \mathbb{E}_{y\sim\mu}\left[ R(y) \sum_{t=1}^{T} \big(\rho_t - 1\big)\Gamma_{t+1} \right],$$

where $\Gamma_{t+1} := \prod_{j=t+1}^{T} \rho_j$.

For each $t$, define $m_t := \min\{t + N - 1, T\}$. Then the full forward trace decomposes as

$$\Gamma_{t+1} = \Gamma_{t+1}^{(N)} Q_{m_t+1},$$

where $Q_{m_t+1} := \prod_{j=m_t+1}^{T} \rho_j$. When $m_t = T$, the residual product $Q_{m_t+1}$ is an empty product and equals one.

Therefore, we have

$$\begin{aligned}
\mathcal{J}(\pi) - \mathcal{J}(\mu) - \mathcal{L}_\mu^{(N)}(\pi) &= \mathbb{E}_{y\sim\mu}\left[ R(y) \sum_{t=1}^{T} \big(\rho_t - 1\big) \left(\Gamma_{t+1} - \Gamma_{t+1}^{(N)}\right) \right] \\
&= \mathbb{E}_{y\sim\mu}\left[ R(y) \sum_{t=1}^{T} \big(\rho_t - 1\big)\Gamma_{t+1}^{(N)} \left(Q_{m_t+1} - 1\right) \right] \\
&\leqslant \xi \sum_{t=1}^{T} \mathbb{E}_\mu\left[ |\rho_t - 1|\, \Gamma_{t+1}^{(N)}\, |Q_{m_t+1} - 1| \right], \qquad \text{(B.7)}
\end{aligned}$$

where the last inequality follows from the assumption $|R(y)| \leqslant \xi$.

We now bound each summand. First, for any state $s$, we get

$$\begin{aligned}\mathbb{E}_{a\sim\mu(\cdot|s)}\left[\left|\frac{\pi(a|s)}{\mu(a|s)}-1\right|\right] &= \sum_a \mu(a|s)\left|\frac{\pi(a|s)}{\mu(a|s)}-1\right| = 2D_{\mathrm{TV}}\left(\mu(\cdot|s),\pi(\cdot|s)\right)\\ &\leqslant 2D_{\mathrm{TV}}^{\max}(\mu,\pi).\end{aligned}$$

For brevity, we write $\delta := D_{\mathrm{TV}}^{\max}(\mu,\pi)$. Thus, we have

$$\mathbb{E}_{y_t\sim\mu(\cdot|s_t)}\left[|\rho_t - 1| \mid s_t\right] \leqslant 2\delta. \tag{B.8}$$

Second, the residual future-token likelihood ratio $Q_{m_t+1} = \prod_{j=m_t+1}^{T}\rho_j$ is the likelihood ratio of the remaining tokens $y_{m_t+1:T}$ under $\pi$ relative to $\mu$, conditional on the prefix up to token $m_t$. Let $P_\mu^{T-m_t}(\cdot|s_{m_t+1})$ and $P_\pi^{T-m_t}(\cdot|s_{m_t+1})$ denote the conditional distributions of the future token sequence $y_{m_t+1:T}$ under $\mu$ and $\pi$, respectively. Then, conditional on $s_{m_t+1}$, we have

$$\begin{aligned}\mathbb{E}_\mu\left[|Q_{m_t+1}-1| \,|\, s_{m_t+1}\right] &= \mathbb{E}_\mu\left[\left|\frac{\prod_{j=m_t+1}^{T}\pi(y_j|s_j)}{\prod_{j=m_t+1}^{T}\mu(y_j|s_j)}-1\right| \,\middle|\, s_{m_t+1}\right]\\ &= \sum_{y_{m_t+1:T}}\left(\prod_{j=m_t+1}^{T}\mu(y_j|s_j)\right)\left|\frac{\prod_{j=m_t+1}^{T}\pi(y_j|s_j)}{\prod_{j=m_t+1}^{T}\mu(y_j|s_j)}-1\right|\\ &= \sum_{y_{m_t+1:T}}\left|\prod_{j=m_t+1}^{T}\pi(y_j|s_j) - \prod_{j=m_t+1}^{T}\mu(y_j|s_j)\right|\\ &= 2D_{\mathrm{TV}}\left(P_\mu^{T-m_t}(\cdot|s_{m_t+1}), P_\pi^{T-m_t}(\cdot|s_{m_t+1})\right),\end{aligned}$$

where last equality follows from $D_{\mathrm{TV}}(P,Q) := \frac{1}{2}\sum_y |P(y)-Q(y)|$.

By the standard telescoping bound for total variation over sequential kernels, we have

$$D_{\mathrm{TV}}\left(P_\mu^{T-m_t}(\cdot|s_{m_t+1}), P_\pi^{T-m_t}(\cdot|s_{m_t+1})\right) \leqslant (T-m_t)\delta.$$

Consequently, we obtain

$$\mathbb{E}_\mu\left[|Q_{m_t+1}-1| \,|\, s_{m_t+1}\right] \leqslant 2(T-m_t)\delta. \tag{B.9}$$

Plugging (B.9) into (B.7), we have

$$\begin{aligned}\mathbb{E}_\mu\left[|\rho_t(\theta)-1|\,\Gamma_{t+1}^{(N)}\,|Q_{m_t+1}-1|\right] &= \mathbb{E}_\mu\left[|\rho_t(\theta)-1|\,\Gamma_{t+1}^{(N)}\mathbb{E}_\mu\left[|Q_{m_t+1}-1| \,|\, s_{m_t+1}\right]\right]\\ &\leqslant 2(T-m_t)\delta\,\mathbb{E}_\mu\left[|\rho_t-1|\,\Gamma_{t+1}^{(N)}\right].\end{aligned} \tag{B.10}$$

Now condition on $s_t$. Since $\Gamma_{t+1}^{(N)}$ is the likelihood-ratio product over the future window $t+1,\ldots,m_t$, its conditional expectation under $\mu$, given $s_{t+1}$, is one, i.e.,

$$\mathbb{E}_\mu\left[\Gamma_{t+1}^{(N)} \,\middle|\, s_{t+1}\right] = 1.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_\mu\left[|\rho_t-1|\,\Gamma_{t+1}^{(N)} \,\middle|\, s_t\right] &= \mathbb{E}_{y_t\sim\mu(\cdot|s_t)}\left[|\rho_t-1|\,\mathbb{E}_\mu\left[\Gamma_{t+1}^{(N)} \,\middle|\, s_{t+1}\right] \,\middle|\, s_t\right]\\ &= \mathbb{E}_{y_t\sim\mu(\cdot|s_t)}\left[|\rho_t-1| \mid s_t\right]\\ &\leqslant 2\delta. \qquad \text{(Eqn. (B.8))}\end{aligned}$$

Combining this bound with (B.10), we get

$$\mathbb{E}_\mu\left[|\rho_t-1|\,\Gamma_{t+1}^{(N)}\,|Q_{m_t+1}-1|\right] \leqslant 4(T-m_t)\delta^2.$$

Summing over $t$, we obtain

$$\begin{aligned}\left|\mathcal{J}(\pi)-\mathcal{J}(\mu)-\mathcal{L}_{\mu}^{(N)}(\pi)\right| &\leqslant 4\xi\delta^2\sum_{t=1}^{T}(T-m_t) = 4\xi\delta^2\sum_{t=1}^{T}(T-t-N+1)_+\\ &= 4\xi\delta^2\sum_{t=1}^{T-N}(T-N+1-t)\\ &= 2\xi(T-N)(T-N+1)\delta^2.\end{aligned}$$

This implies that

$$\mathcal{J}(\pi)-\mathcal{J}(\mu)\geqslant\mathcal{L}_{\mu}^{(N)}(\pi)-2\xi(T-N)(T-N+1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2.$$

The second term on the right-hand side is the bias upper bound, whose magnitude is strictly decreasing in $N$. This proves the first inequality.

**Step 2: Empirical concentration.** Define the single-sample $N$-step surrogate estimator

$$Z_N(y):=R(y)\sum_{t=1}^{T}\big(\rho_t-1\big)\Gamma_{t+1}^{(N)}.$$

Then the population $N$-step surrogate and its empirical estimate can be written as

$$\mathcal{L}_{\mu}^{(N)}(\pi)=\mathbb{E}_{y\sim\mu}[Z_N(y)],\qquad\text{and}\qquad\widehat{\mathcal{L}}_{\mu}^{(N)}(\pi)=\frac{1}{G}\sum_{i=1}^{G}Z_N(y^{(i)}).$$

Let $k_t(N):=\min\{N-1,T-t\}$. The product $\Gamma_{t+1}^{(N)}$ contains exactly $k_t(N)$ future likelihood-ratio factors. Since $|\rho_t-1|\leqslant\epsilon$, we have $0\leqslant\rho_t\leqslant1+\epsilon$. Therefore, we get

$$0\leqslant\Gamma_{t+1}^{(N)}\leqslant(1+\epsilon)^{k_t(N)}$$

Using $|R(y)|\leqslant\xi$, we obtain

$$\begin{aligned}|Z_N(y)| &= \left|R(y)\sum_{t=1}^{T}\big(\rho_t-1\big)\Gamma_{t+1}^{(N)}\right|\\ &\leqslant |R(y)|\sum_{t=1}^{T}|\rho_t-1|\,\Gamma_{t+1}^{(N)}\\ &\leqslant \xi\epsilon\sum_{t=1}^{T}(1+\epsilon)^{\min\{N-1,T-t\}}=:B_N.\end{aligned}$$

Thus, $Z_N(y)\in[-B_N,B_N]$.

Since $y^{(1)},\ldots,y^{(G)}$ are i.i.d. samples from $\mu$, the random variables $Z_N(y^{(1)}),\ldots,Z_N(y^{(G)})$ are also i.i.d. and bounded. By Hoeffding's inequality, for any $a>0$, we get

$$\mathbb{P}\left(\widehat{\mathcal{L}}_{\mu}^{(N)}(\pi)-\mathcal{L}_{\mu}^{(N)}(\pi)\geqslant a\right)\leqslant\exp\left(-\frac{Ga^2}{2B_N^2}\right).$$

Taking $a=B_N\sqrt{\frac{2\log(1/\alpha)}{G}}$, we obtain that, with probability at least $1-\alpha$,

$$\mathcal{L}_{\mu}^{(N)}(\pi)\geqslant\widehat{\mathcal{L}}_{\mu}^{(N)}(\pi)-B_N\sqrt{\frac{2\log(1/\alpha)}{G}}.$$

Combining this with the population truncation bound from Step 1 gives

$$\mathcal{J}(\pi)-\mathcal{J}(\mu)\geqslant\widehat{\mathcal{L}}_{\mu}^{(N)}(\pi)-2\xi(T-N)(T-N+1)\Big(D_{\mathrm{TV}}^{\max}(\mu,\pi)\Big)^2-B_N\sqrt{\frac{2\log(1/\alpha)}{G}}.$$

This proves the high-probability empirical policy-improvement bound.

**Step 3: Monotonicity of the concentration penalty.** It remains to show that $B_N$ is strictly increasing in $N$. Define $S_N := \sum_{t=1}^{T}(1+\epsilon)^{\min\{N-1,T-t\}}$, so that $B_N = \xi\epsilon S_N$. Then, for $N = 1, \ldots, T-1$, we have

$$S_{N+1} - S_N = \sum_{t=1}^{T} \left( (1+\epsilon)^{\min\{N,T-t\}} - (1+\epsilon)^{\min\{N-1,T-t\}} \right).$$

If $t > T - N$, then both minima are equal to $T - t$, so the corresponding summand is zero. If $t \leqslant T - N$, then $\min\{N, T-t\} = N$ and $\min\{N-1, T-t\} = N-1$. Therefore, we obtain

$$\begin{aligned} S_{N+1} - S_N &= \sum_{t=1}^{T-N} \left( (1+\epsilon)^N - (1+\epsilon)^{N-1} \right) \\ &= (T-N)\left( (1+\epsilon)^N - (1+\epsilon)^{N-1} \right) \\ &= (T-N)(1+\epsilon)^{N-1}\epsilon > 0. \end{aligned}$$

Thus $S_N$, and therefore $B_N = \xi\epsilon S_N$, is strictly increasing in $N$. Consequently, the Hoeffding concentration penalty $B_N\sqrt{\frac{2\log(1/\alpha)}{G}}$ is also strictly increasing in $N$. This completes the proof of Theorem 1. □

# C Implementation Details

**Framework.** We implemented NFPO using the verl [34] framework, which is a widely used training pipeline for LLM reinforcement learning. The experiments with baseline algorithms follow the official implementation included in the verl framework, and we additionally implemented the loss function for NFPO. Since other components of the training pipeline, including data generation, reward computation, and advantage estimation, remain unchanged, our experiments are easily reproducible by replacing the loss function. All models are trained on 8 H200 GPUs. Each training run requires approximately one day for the 1.7B model and two days for the 8B model.

**Hyperparameters.** Unless otherwise specified, the hyperparameters for the main mathematical reasoning experiments follow the default settings of verl. Table C.1 summarizes the hyperparameters shared across all algorithms in these experiments.

For GRPO, we use the standard clip-higher configuration ($\epsilon_{\text{low}} = 0.2$, $\epsilon_{\text{high}} = 0.28$). For DPPO and NFPO, we use a total variation–based token mask with threshold $\delta = 0.2$, following Qi et al. [28]. We do not apply auxiliary KL or entropy regularization.

For the main mathematical reasoning experiments, NFPO uses a forward-trace clipping range of $[0.8, 1.4]$ and a likelihood-ratio clipping threshold of $\beta = 3.0$. We set the trace horizon to $N = 4$ for Qwen3-1.7B-Base and $N = 8$ for Qwen3-8B-Base.

For the additional-domain experiments in Section E.2, we use the same configuration across all four domains, with $N = 4$ and a forward-trace clipping range of $[0.8, 1.2]$. Countdown follows the LPO configuration on the Countdown Number Game [29, 25]. PRIME Code uses the code split of `PRIME-RL/Eurus-2-RL-Data` and the corresponding `prime_code` reward [5]. ALFWorld follows the GIGPO configuration [35, 7]. Sudoku follows the DPPO configuration on the GEM environment [28, 19]. We use Qwen3-1.7B-Base for Countdown, PRIME Code, and Sudoku, and Qwen2.5-1.5B-Instruct for ALFWorld.

For the token-mask ablation in Figure 5, we use the standard hyperparameters for each mask. We use PPO clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.28$, the IcePop mask with ratio threshold $\beta = 3.0$, and the KL-divergence mask with threshold $\delta = 0.05$.

**Details on the Token MDP in Section 3.** We consider a simple token-level MDP defined over a fixed vocabulary $\mathcal{V} = a, b, c$ with a finite horizon $T = 7$. A trajectory corresponds to a sequence $y = (y_1, \ldots, y_T) \in \mathcal{V}^T$, which can be interpreted as a string generated autoregressively. At each timestep $t$, the state is given by the prefix $s_t = (y_1, \ldots, y_{t-1})$, and the policy outputs a distribution over the next token.

Table C.1: Common hyperparameters shared across all algorithms.

| Category | Hyperparameter | Value |
|---|---|---|
| Optimization | Learning Rate | $1 \times 10^{-6}$ |
| | Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$) |
| | Weight Decay | 0.01 |
| Data | Prompt per Batch | 128 |
| | Rollouts per Prompt ($G$) | 8 |
| | Mini-batch Size | 256 (across all DP ranks) |
| | Rollout Temperature | 1.0 for training, 0.7 for evaluation |
| | Max Response Tokens | 8,000 (MATH), 16,000 (DAPO-Math-17k) |
| | Training Steps | 500 (MATH), 300 (DAPO-Math-17k) |
| Algorithm | KL Loss | No |
| | Entropy Loss | No |
| | Log probability recompute | No |

The reward is defined as a sparse binary signal based on whether the generated sequence contains a predefined target pattern as a subsequence. Concretely, given a target string $y^\star = \texttt{abcabc}$, the reward function is

$$R(y) := \mathbf{1}\{y^\star \text{ is a subsequence of } y\}$$

i.e., the reward is 1 if the sequence $y$ contains $y^\star$ in order (not necessarily contiguously), and 0 otherwise. This reward structure introduces a long-horizon dependency, as achieving a nonzero reward requires consistently selecting tokens that match the target pattern throughout the sequence.

Both the rollout policy $\mu$ and the target policy $\pi$ are instantiated from the same parametric family, which we refer to as the *target-following policy*. Given a prefix $s$, let $k(s)$ denote the length of the longest prefix of $y^\star$ that appears as a subsequence in $s$. The policy then assigns probability $\alpha$ to the next required target token $y^\star_{k(s)+1}$, and distributes the remaining mass uniformly over the other tokens:

$$\pi_\alpha(v|s) = \begin{cases} \alpha & \text{if } v = y^\star_{k(s)+1} \\ \frac{1-\alpha}{|\mathcal{V}|-1} & \text{otherwise.} \end{cases}$$

When the full target has already been matched ($k(s) = |y^\star|$), the policy becomes uniform over the vocabulary.

We construct $\mu$ and $\pi$ by choosing different values of $\alpha$, with $\alpha_\mu < \alpha_\pi$. Intuitively, both policies are biased toward completing the target pattern, but $\pi$ is more strongly aligned with the optimal behavior. In our experiments, we use $\alpha_\mu = 0.5$ and $\alpha_\pi = 0.8$.

# D Detailed Token Analysis

This appendix expands the analysis in Section 5 along three directions. Section D.1 gives the rule-based definitions of the token categories used throughout the analysis. Section D.2 reports the full per-category correction strength of token-level weights under GRPO and NFPO that is summarized in the main section. Section D.3 presents representative rollout cases with token-level tables of the likelihood ratio $\rho_t$, the raw forward trace $\Gamma^{(N)}_{t+1}$, the clipped trace $\bar{\Gamma}^{(N)}_{t+1}$, and the token advantage; orange highlights mark spans whose forward trace is clipped at either the upper or lower bound, and the highlight is meant to identify a local span of clipped tokens.

## D.1 Token Categories

Every vocabulary token is assigned to exactly one of eight categories by a rule-based classifier on the token string.

- **Reasoning connectives.** Logical and discourse connectors that link reasoning steps within or across sentences, e.g. *therefore, since, because, however, thus, also, first, next, finally, wait, actually*.

- **Reasoning actions.** Verbs, modals, and pronouns that introduce a reasoning step or operation, e.g. *we, let, can, need, find, solve, use, consider, assume, check, verify, suppose, given, define*.
- **Function words.** Closed-class English function words: articles, prepositions, pronouns, and auxiliaries (*the, of, to, is, and, that, which*).
- **Math content nouns.** Mathematical content nouns and adjectives such as *equation, function, variable, root, factor, triangle, integer, prime, solution*.
- **Math variables.** Single-letter variable tokens such as *x, y, n, k, b*, excluding single-letter English words *a*, *A*, and *I*.
- **Numeric.** Digits, decimals, arithmetic and comparison operators, LaTeX math commands such as `\frac`, `\sqrt`, `\cdot`, and number words (*one, two, hundred, half, quarter*).
- **Formatting.** Whitespace, BPE prefix-only tokens, newlines, brackets, punctuation runs, markdown markers, special tokens, and LaTeX formatting commands such as `\boxed`, `\begin`, `\text`, `\left`, `\right`.
- **Other.** Residual BPE pieces and rare strings that do not match any of the above.

Table D.1: Per-category correction strength of token-level weights.

| | Correction strength | |
|---|---|---|
| Category | $\rho_t$ | $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$ |
| All tokens (baseline) | 0.02775 | 0.07218 |
| Reasoning action | 0.05167 | 0.11118 |
| Reasoning connective | 0.05595 | 0.10099 |
| Function word | 0.03834 | 0.09618 |
| Math content noun | 0.04701 | 0.10218 |
| Other (residual) | 0.04025 | 0.09044 |
| Math variable (single letter) | 0.02387 | 0.06332 |
| Formatting | 0.02214 | 0.06603 |
| Numeric | 0.01789 | 0.05477 |

### D.2 Correction Strength of Per-Token Weights by Category

This subsection reports the full per-category correction strength. The pool consists of approximately $3.4$M valid response tokens sampled during NFPO training. For each category, Table D.1 reports the correction strength of $\rho_t$ and $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$ for NFPO.

Two patterns are consistent across categories. First, $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$ achieves substantially larger correction strength than $\rho_t$ in every category, confirming that the forward trace produces more active corrections than the likelihood ratio alone. Second, reasoning-related categories (reasoning action, reasoning connective) and content-bearing categories (function word, math content noun) carry substantially larger correction strength than residual categories such as numeric, formatting, and single-letter math variables under both $\rho_t$ and $\rho_t \cdot \bar{\Gamma}_{t+1}^{(N,i)}$. This pattern supports the claim that the forward trace amplifies corrections on the tokens that drive reasoning behavior.

### D.3 Case study

This subsection presents representative cases from rollouts during NFPO training. The first four cases show situations where the forward trace amplifies the update, and the model is rewarded more strongly than the local ratio alone would suggest. These spans typically initiate or transition between reasoning steps, and the forward trace pushes the policy further in their direction. The last case shows the opposite situation, where the forward trace dampens the update on a failed trajectory. Even when the final answer is wrong, spans that contain valid reasoning attempts receive a softer penalty, so that the update does not punish the reasoning move itself.

**Case 1: Goal definition**

**Prompt.** Palindromic primes are two-digit prime numbers whose reversed digits also form a prime. Find the sum of all palindromic primes less than $50$.

**Response excerpt.**
To find the sum of all palindromic primes less than $50$, `we first need to identify all two-digit prime numbers that` remain prime when their digits are reversed. ...

| idx | token | $\rho_t$ | $\Gamma^{(N)}_{t+1}$ | $\bar{\Gamma}^{(N)}_{t+1}$ | $\hat{A}_t$ |
|---|---|---|---|---|---|
| 17 | `we` | 1.000 | 2.67 | 1.400 | +0.625 |
| 18 | `first` | 0.976 | 2.74 | 1.400 | +0.625 |
| 19 | `need` | 1.000 | 2.74 | 1.400 | +0.625 |
| 20 | `to` | 1.000 | 2.77 | 1.400 | +0.625 |
| 21 | `identify` | 1.000 | 4.77 | 1.400 | +0.625 |
| 22 | `all` | 2.737 | 1.74 | 1.400 | +0.625 |
| 23 | `two` | 1.000 | 1.74 | 1.400 | +0.625 |
| 24 | `-digit` | 1.000 | 1.74 | 1.400 | +0.625 |
| 25 | `prime` | 1.000 | 1.74 | 1.400 | +0.625 |
| 26 | `numbers` | 1.000 | 1.74 | 1.400 | +0.625 |
| 27 | `that` | 1.011 | 1.73 | 1.400 | +0.625 |
| 28 | `remain` | 1.725 | 1.00 | 1.000 | +0.625 |

The highlighted span is an early planning phrase, not part of the final answer. Several tokens have local ratios close to one, but their forward traces are large enough to hit the upper clipping bound because the continuation opened by this subgoal becomes substantially more likely under the updated policy. This shows that the forward trace captures a span-level signal that is mostly invisible to the local token ratio alone.

**Case 2: Self-correction to an alternative approach**

**Prompt.** Each face of a cube is painted either red or blue, each with probability $1/2$. The color of each face is determined independently. What is the probability that the painted cube can be placed on a horizontal surface so that the four vertical faces are all the same color?

**Response excerpt.**
we can find a smarter way.

Considering the Complementary Event:
Let's think about the complementary event: it is counted `4 times in total . Wait , let 's think differently .`
Total colorings: $2^6 = \cdots$.

| idx | token | $\rho_t$ | $\Gamma^{(N)}_{t+1}$ | $\bar{\Gamma}^{(N)}_{t+1}$ | $\hat{A}_t$ |
|---|---|---|---|---|---|
| 2860 | `·is` | 0.965 | 0.97 | 0.973 | +0.875 |
| 2861 | `·counted` | 1.006 | 0.97 | 0.965 | +0.875 |
| 2862 | `·` | 1.000 | 0.96 | 0.965 | +0.875 |
| 2863 | `4` | 0.999 | 1.58 | 1.400 | +0.875 |
| 2864 | `·times` | 1.000 | 1.58 | 1.400 | +0.875 |
| 2865 | `·in` | 1.000 | 1.59 | 1.400 | +0.875 |
| 2866 | `·total` | 1.002 | 1.82 | 1.400 | +0.875 |
| 2867 | `.\n\n` | 0.965 | 1.88 | 1.400 | +0.875 |
| 2868 | `Wait` | **0.998** | **2.94** | **1.400** | +0.875 |
| 2869 | `,` | 1.000 | 2.95 | 1.400 | +0.875 |
| 2870 | `·let` | 1.637 | 1.80 | 1.400 | +0.875 |
| 2871 | `'s` | 1.000 | 1.80 | 1.400 | +0.875 |
| 2872 | `·think` | 1.006 | 1.79 | 1.400 | +0.875 |
| 2873 | `·differently` | 1.147 | 1.56 | 1.400 | +0.875 |
| 2874 | `.\n\n` | 1.000 | 1.56 | 1.400 | +0.875 |
| 2875 | `Total` | 1.559 | 1.00 | 1.003 | +0.875 |
| 2876 | `·color` | 1.003 | 1.00 | 1.000 | +0.875 |
| 2877 | `ings` | 1.000 | 1.00 | 1.000 | +0.875 |
| 2878 | `:` | 1.000 | 1.00 | 1.000 | +0.875 |

| | | | | | |
|---|---|---|---|---|---|
| 2879 | ·\( | 1.000 | 1.00 | 1.000 | +0.875 |
| 2880 | 2 | 1.000 | 1.00 | 1.000 | +0.875 |
| 2881 | ^ | 1.000 | 1.00 | 1.000 | +0.875 |
| 2882 | 6 | 1.000 | 1.04 | 1.044 | +0.875 |
| 2883 | ·= | 1.000 | 1.04 | 1.044 | +0.875 |
| 2884 | · | 1.000 | 1.04 | 1.038 | +0.875 |

This case shows a clear self-correction in the reasoning trajectory. The model first attempts a complementary-event argument, notices that the counting structure is becoming awkward, and then switches strategy with the phrase "Wait, let's think differently." The focus token `Wait` has an almost unchanged local ratio ($\rho_t = 0.998$), but its raw forward trace rises to $\Gamma^{(N)}_{t+1} = 2.94$ and is clipped to $\bar{\Gamma}^{(N)}_{t+1} = 1.400$. This is a case where the local ratio fails to capture but the forward trace captures.

**Case 3: Causal connector**

**Prompt.** Find the remainder when $1 + 2 + 2^2 + \cdots + 2^{100}$ is divided by 7.

**Response excerpt.**
$\ldots$ the exponent modulo 3. `Since the pattern repeats every 3` exponents, we can find $2^{101} \bmod 7$. $\ldots$

| idx | token | $\rho_t$ | $\Gamma^{(N)}_{t+1}$ | $\bar{\Gamma}^{(N)}_{t+1}$ | $\hat{A}_t$ |
|---|---|---|---|---|---|
| 636 | Since | 1.000 | 2.66 | 1.400 | +0.875 |
| 637 | the | 1.000 | 2.66 | 1.400 | +0.875 |
| 638 | pattern | 1.015 | 2.62 | 1.400 | +0.875 |
| 639 | repeats | 1.000 | 2.62 | 1.400 | +0.875 |
| 640 | every | 1.000 | 2.62 | 1.400 | +0.875 |
| 641 | · | 1.000 | 2.62 | 1.400 | +0.875 |
| 642 | 3 | 1.000 | 2.62 | 1.400 | +0.875 |
| 643 | ex | 2.618 | 1.00 | 0.999 | +0.875 |
| 644 | ponents | 1.000 | 0.97 | 0.967 | +0.875 |

The highlighted span is a causal transition into the periodicity argument. The tokens themselves do not perform the computation, but together they mark the step where a modular pattern becomes the final answer. The forward trace assigns high weight to this connector span, even though the local ratios alone leave it nearly unchanged.

**Case 4: Consolidation connector**

**Prompt.** A standard six-sided die is rolled twice. What is the probability that the difference between the two rolls is 2?

**Response excerpt.**
$\ldots$ $y = 4, 8$) `( Invalid ) So , the valid` pairs are: $(1, 3), (2, 4), \ldots$

| idx | token | $\rho_t$ | $\Gamma^{(N)}_{t+1}$ | $\bar{\Gamma}^{(N)}_{t+1}$ | $\hat{A}_t$ |
|---|---|---|---|---|---|
| 1439 | ) | 1.000 | 2.95 | 1.400 | +0.375 |
| 1440 | ( | 1.000 | 4.14 | 1.400 | +0.375 |
| 1441 | Invalid | 1.000 | 4.14 | 1.400 | +0.375 |
| 1442 | )\n\n | 1.000 | 4.14 | 1.400 | +0.375 |
| 1443 | So | 1.000 | 4.14 | 1.400 | +0.375 |
| 1444 | , | 1.000 | 4.14 | 1.400 | +0.375 |
| 1445 | the | 1.000 | 4.14 | 1.400 | +0.375 |
| 1446 | valid | 2.954 | 1.40 | 1.400 | +0.375 |
| 1447 | pairs | 1.404 | 1.00 | 1.000 | +0.375 |
| 1448 | are | 1.000 | 1.00 | 1.000 | +0.375 |
| 1449 | : | 1.000 | 1.00 | 1.000 | +0.375 |

The highlighted span is a consolidation connector that moves from rejecting invalid outcomes to listing the valid pairs. The forward trace weights this consolidation step as a whole rather than to any single token.

**Case 5: Self-correction in a failed trajectory**

**Prompt.** Either increasing the radius or the height of a cylinder by six inches will result in the same volume. The original height of the cylinder is two inches. What is the original radius in inches?

**Response excerpt.**
$\ldots\ 2\sqrt{7}\,]$ `However , this does not fit the` context of the problem. Let's re-evaluate $\ldots$

| idx | token | $\rho_t$ | $\Gamma^{(N)}_{t+1}$ | $\bar{\Gamma}^{(N)}_{t+1}$ | $\hat{A}_t$ |
|---|---|---|---|---|---|
| 867 | `{` | 1.000 | 0.90 | 0.900 | $-0.500$ |
| 868 | `7` | 1.000 | 0.70 | 0.800 | $-0.500$ |
| 869 | `}` | 1.000 | 0.70 | 0.800 | $-0.500$ |
| 870 | `·\` | 1.000 | 0.70 | 0.800 | $-0.500$ |
| 871 | `]\n\n` | 1.000 | 0.70 | 0.800 | $-0.500$ |
| 872 | `However` | **0.998** | **0.56** | **0.800** | $-0.500$ |
| 873 | `,` | 1.000 | 0.56 | 0.800 | $-0.500$ |
| 874 | `·this` | 0.901 | 0.62 | 0.800 | $-0.500$ |
| 875 | `·does` | 0.782 | 0.80 | 0.800 | $-0.500$ |
| 876 | `·not` | 1.000 | 0.80 | 0.800 | $-0.500$ |
| 877 | `·fit` | 1.000 | 0.80 | 0.800 | $-0.500$ |
| 878 | `·the` | 0.999 | 0.80 | 0.800 | $-0.500$ |
| 879 | `·context` | 0.798 | 0.98 | 0.984 | $-0.500$ |
| 880 | `·of` | 0.999 | 1.05 | 1.047 | $-0.500$ |
| 881 | `·the` | 1.000 | 1.05 | 1.047 | $-0.500$ |
| 882 | `·problem` | 1.000 | 1.05 | 1.047 | $-0.500$ |
| 883 | `.` | 0.999 | 1.08 | 1.083 | $-0.500$ |
| 884 | `·Let` | 1.000 | 1.05 | 1.052 | $-0.500$ |
| 885 | `'s` | 1.000 | 1.05 | 1.052 | $-0.500$ |
| 886 | `·re` | 0.986 | 1.05 | 1.052 | $-0.500$ |
| 887 | `-e` | 1.063 | 0.94 | 0.942 | $-0.500$ |
| 888 | `valuate` | 1.000 | 0.94 | 0.942 | $-0.500$ |

This case is the symmetric counterpart of Case 2. The trajectory ultimately fails (reward 0) and the global advantage is therefore negative ($\hat{A}_t = -0.500$). A purely local surrogate would penalize every token in the trajectory by the same advantage scale. The forward trace, however, drops to $\Gamma^{(N)}_{t+1} = 0.56$ on the token `However` and stays low across the whole self-correction phrase "However, this does not fit the context ... Let's re-evaluate"; the clipped trace at the lower bound $0.8$ still reduces the effective gradient on this span by $20\%$ relative to the local-only update.

# E More Experimental Results

This section provides additional experimental results supporting the main evaluation in Section 4 and the analyses in Section 6. We organize the results into four parts. First, we provide detailed learning dynamics and additional evaluations on the main mathematical reasoning benchmarks, including confidence intervals and pass@$k$ results. Second, we provide the experimental settings and validation curves for the additional domains. Third, we provide additional robustness and ablation studies across trace horizons, response-length regimes, clipping ranges, and other design choices. Fourth, we report the computational and memory overhead of NFPO compared to GRPO and DPPO.

## E.1 Extended Results on the Main Mathematical Reasoning Benchmarks

**Detailed learning dynamics.** Figures E.1 and E.2 compare the training dynamics of GRPO, GSPO, DPPO, and NFPO on Qwen3-8B-Base and Qwen3-1.7B-Base, respectively. We report the training reward, mean response length, KL divergence, and per-benchmark accuracy throughout training.

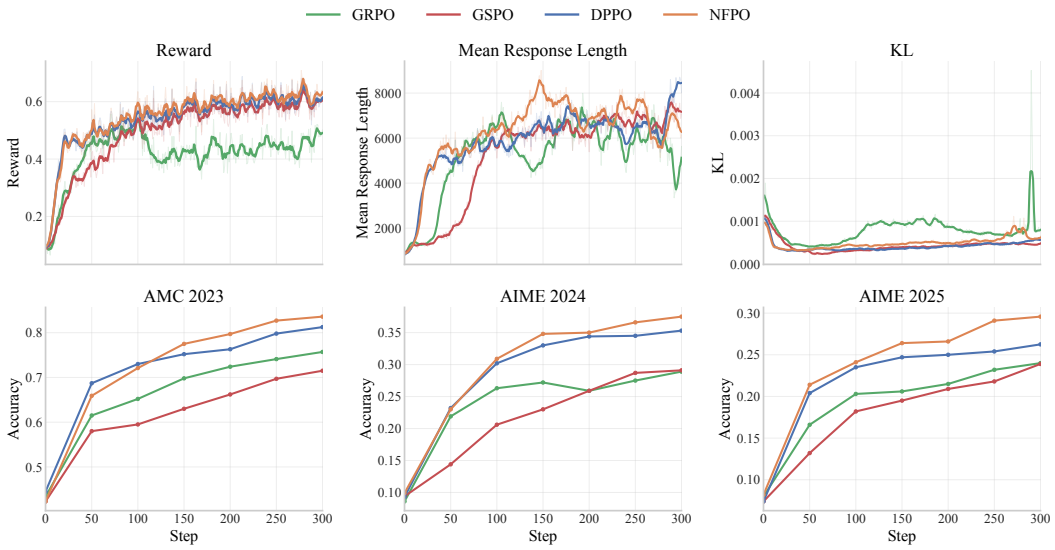



Figure E.1: Learning dynamics on Qwen3-8B-Base. Top row: training reward, mean response length, and KL divergence. Bottom row: per-benchmark accuracy.

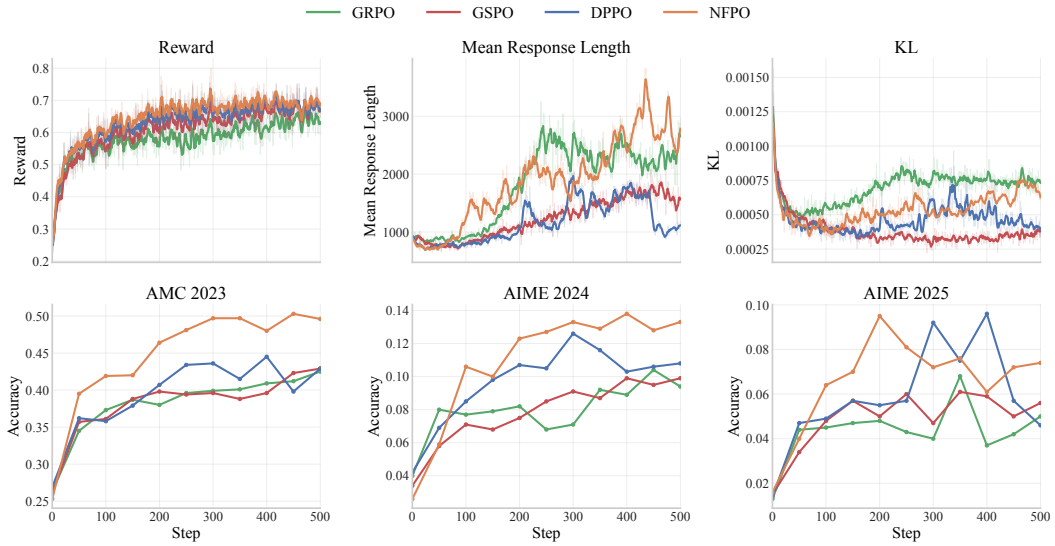



Figure E.2: Learning dynamics on Qwen3-1.7B-Base. Top row: training reward, mean response length, and KL divergence. Bottom row: per-benchmark accuracy.

**Confidence interval analysis.** NFPO shows a consistent and statistically supported advantage on AIME: it achieves higher point estimates than GRPO, GSPO and DPPO in all six comparisons, with non-overlapping 95% confidence intervals in all comparisons against GRPO and two comparisons against DPPO. The intervals overlap with those of DPPO in the remaining four comparisons. Table E.1 reports the full results.

Each reported pass@1 score is averaged over 32 sampled responses per prompt at each of the final three evaluation checkpoints, resulting in $30 \times 32 \times 3 = 2{,}880$ scored responses per benchmark. We compute Wald confidence intervals by treating the scored responses as independent.

Table E.1: AIME pass@1 with 95% Wald confidence intervals.

| Model | Method | AIME24 | AIME25 | AIME26 |
|---|---|---|---|---|
| Qwen3-1.7B-Base | Base | 0.034 [0.027, 0.041] | 0.015 [0.011, 0.019] | 0.021 [0.016, 0.026] |
| | GRPO | 0.094 [0.083, 0.105] | 0.050 [0.042, 0.058] | 0.047 [0.039, 0.055] |
| | GSPO | 0.098 [0.087, 0.108] | 0.055 [0.047, 0.064] | 0.053 [0.045, 0.061] |
| | DPPO | 0.106 [0.095, 0.117] | 0.066 [0.057, 0.075] | 0.056 [0.048, 0.065] |
| | NFPO | **0.133** [0.121, 0.146] | **0.074** [0.064, 0.084] | **0.081** [0.071, 0.091] |
| Qwen3-8B-Base | Base | 0.094 [0.083, 0.105] | 0.074 [0.064, 0.084] | 0.056 [0.048, 0.064] |
| | GRPO | 0.274 [0.258, 0.290] | 0.229 [0.214, 0.245] | 0.230 [0.215, 0.245] |
| | GSPO | 0.291 [0.262, 0.319] | 0.239 [0.212, 0.266] | 0.258 [0.231, 0.286] |
| | DPPO | 0.347 [0.330, 0.365] | 0.256 [0.240, 0.272] | 0.265 [0.249, 0.281] |
| | NFPO | **0.364** [0.346, 0.381] | **0.284** [0.268, 0.301] | **0.286** [0.269, 0.302] |

**Pass@$k$ evaluation.** NFPO improves pass@32 at both model scales, suggesting that its performance gains do not come at the expense of coverage over successful solutions. It achieves the highest benchmark-averaged pass@32, improving over DPPO from 0.496 to 0.518 on Qwen3-1.7B-Base and from 0.696 to 0.734 on Qwen3-8B-Base. It also outperforms DPPO in all 14 benchmark–model comparisons. Table E.2 reports the full results.

Large-$k$ pass@$k$ is not a direct measure of diversity, but it is commonly used as a proxy for coverage over successful solutions and exploration across reasoning paths [39, 15, 46]. A decrease in large-$k$ pass@$k$ together with an increase in pass@1 can indicate that the policy has concentrated on a narrower set of solutions.

Table E.2: Pass@32 on mathematical reasoning benchmarks.

| Model | Method | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3-1.7B-Base | GRPO | 0.329 | 0.214 | 0.192 | **0.813** | 0.884 | 0.459 | 0.565 | 0.494 |
| | GSPO | 0.296 | 0.204 | 0.189 | 0.788 | 0.888 | 0.421 | 0.577 | 0.480 |
| | DPPO | 0.313 | 0.211 | 0.242 | 0.784 | 0.897 | 0.446 | 0.578 | 0.496 |
| | NFPO | **0.331** | **0.227** | **0.246** | 0.811 | **0.908** | **0.492** | **0.610** | **0.518** |
| Qwen3-8B-Base | GRPO | 0.689 | 0.476 | 0.583 | 0.945 | 0.966 | 0.558 | **0.750** | 0.710 |
| | GSPO | 0.681 | **0.512** | **0.656** | 0.941 | 0.966 | 0.553 | 0.744 | 0.722 |
| | DPPO | 0.655 | 0.414 | 0.586 | 0.949 | 0.962 | 0.572 | 0.734 | 0.696 |
| | NFPO | **0.736** | 0.505 | 0.636 | **0.962** | **0.968** | **0.581** | 0.749 | **0.734** |

## E.2 Extension to Other Domains

This subsection provides the full results and experimental settings for the additional-domain results in Table 2 (Section 4). Table E.3 reports pass@$k$ in addition to pass@1, together with the relative improvements over DPPO, and Figure E.3 shows the corresponding validation curves throughout training.

Table E.3: Full results on four additional domains. Relative improvements are computed with respect to DPPO.

| Domain | Task type | Metric | DPPO | NFPO | Relative gain |
|---|---|---|---|---|---|
| Countdown | Logical reasoning | pass@1 | 0.468 | **0.524** | **+11.9%** |
| | | pass@8 | 0.530 | **0.611** | **+15.2%** |
| PRIME Code | Code generation | pass@1 | 0.236 | **0.263** | **+11.4%** |
| | | pass@4 | 0.283 | **0.321** | **+13.4%** |
| ALFWorld | Multi-turn agentic | pass@1 | 0.765 | **0.860** | **+12.4%** |
| | | pass@8 | 0.875 | **0.938** | **+7.2%** |
| Sudoku | Multi-turn puzzle | Episode return | 0.318 | **0.656** | **+106%** |

Countdown follows the LPO training setup on the Countdown Number Game [29, 25]. PRIME Code uses the code split of `PRIME-RL/Eurus-2-RL-Data` and the corresponding `prime_code` reward [5]. ALFWorld follows the GIGPO configuration on the ALFWorld environment [35, 7]. Sudoku follows the DPPO configuration on the GEM environment [28, 19].

We use Qwen3-1.7B-Base for Countdown, PRIME Code, and Sudoku. For ALFWorld, we use Qwen2.5-1.5B-Instruct following the corresponding training protocol. For Countdown and ALFWorld, the reported pass@1 score is averaged over eight sampled rollouts per input. For PRIME Code, it is averaged over four sampled rollouts per input.

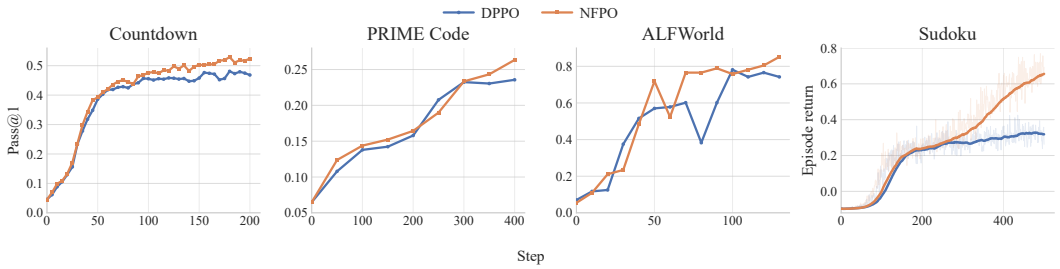



Figure E.3: Validation performance throughout training on Countdown, PRIME Code, ALFWorld, and Sudoku.

### E.3 Ablation Results

This subsection reports the full per-benchmark results for the ablations in Section 6 and presents additional ablations on other domains and model scales.

**Robustness to the trace horizon $N$.** NFPO is robust to the trace horizon across different domains, model scales, and response-length regimes. We evaluate Countdown with Qwen3-1.7B-Base, mathematical reasoning with Qwen3-1.7B-Base, and mathematical reasoning with Qwen3-8B-Base. These settings cover responses of approximately 300–700, 1k–3k, and 6k–8k tokens, respectively.
Across all three settings, moderate trace horizons improve over or closely match the $N = 1$ baseline. Table E.4 reports the pass@1 results on Countdown. Tables E.5 and E.6 report the full per-benchmark results for the two evaluated mathematical reasoning settings.

Table E.4: Trace-horizon ablation on Countdown with Qwen3-1.7B-Base.

| $N$ | **pass@1** |
|---|---|
| 1 (DPPO) | 0.468 |
| 2 | 0.465 |
| **4** | **0.524** |
| 8 | 0.503 |
| 16 | 0.493 |

Table E.5: Effect of the trace horizon $N$ on mathematical reasoning benchmarks using Qwen3-1.7B-Base.

| $N$ | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| 1 (DPPO) | 0.106 | 0.066 | 0.056 | 0.424 | 0.715 | 0.246 | 0.341 | 0.279 |
| 2 | 0.114 | 0.082 | 0.053 | 0.431 | 0.718 | 0.246 | 0.344 | 0.284 |
| 4 | **0.133** | 0.074 | **0.081** | **0.496** | 0.745 | **0.262** | 0.360 | **0.307** |
| 8 | 0.124 | **0.114** | 0.050 | 0.473 | 0.748 | 0.244 | 0.355 | 0.301 |
| 16 | 0.105 | 0.078 | 0.079 | 0.433 | **0.749** | 0.254 | **0.370** | 0.296 |

Table E.6: Effect of the trace horizon $N$ on mathematical reasoning benchmarks using Qwen3-8B-Base.

| $N$ | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| 1 (DPPO) | 0.347 | 0.256 | 0.265 | 0.791 | 0.886 | **0.399** | 0.565 | 0.501 |
| 2 | 0.339 | **0.294** | 0.278 | 0.780 | **0.894** | 0.389 | 0.556 | 0.504 |
| 4 | 0.353 | 0.279 | 0.272 | 0.809 | 0.890 | 0.385 | **0.577** | 0.509 |
| 8 | **0.364** | 0.284 | 0.286 | **0.820** | 0.892 | 0.390 | 0.571 | **0.515** |
| 16 | 0.349 | 0.284 | **0.305** | 0.800 | **0.894** | 0.387 | 0.576 | 0.513 |

**Forward-trace clipping.** The clipping range $[0.8, 1.4]$ achieves the highest average score of $0.307$. Table E.7 reports the full per-benchmark results for $N = 4$.

Table E.7: Effect of the forward-trace clipping range on mathematical reasoning benchmarks with $N = 4$.

| **Clip range** | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| $[0.6, 1.2]$ | 0.117 | 0.067 | 0.050 | **0.504** | 0.737 | **0.270** | 0.359 | 0.300 |
| $[0.8, 1.2]$ | 0.123 | 0.071 | 0.079 | 0.493 | 0.731 | 0.247 | 0.358 | 0.300 |
| $[0.8, 1.4]$ | **0.133** | **0.074** | **0.081** | 0.496 | **0.745** | 0.262 | **0.360** | **0.307** |

**Clipping range across trace horizons.** A tighter clipping range performs better at $N = 4$, while a wider range performs better at $N = 16$ in both mathematical reasoning and Countdown. As shown in Figure E.4, widening the range at $N = 16$ improves the score from $0.296$ to $0.308$ on mathematical reasoning and from $0.493$ to $0.502$ on Countdown. For mathematical reasoning, the wider range at $N = 16$ also slightly exceeds the best $N = 4$ result. This pattern suggests that longer traces benefit from less restrictive clipping, which preserves more of the accumulated correction signal.

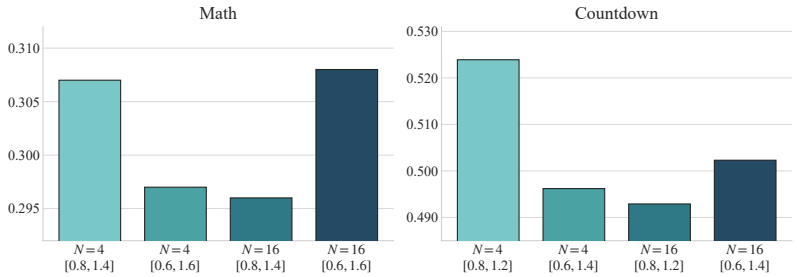



Figure E.4: Effect of the forward-trace clipping range at $N = 4$ and $N = 16$.

**Log-probability recomputation and likelihood-ratio clipping.** The default configuration outperforms log-probability recomputation, while removing token-wise clipping of $\rho_t$ leaves the average score unchanged. Table E.8 reports the full results. Recomputation removes the inference–training mismatch from the likelihood ratio and reduces the average score from $0.307$ to $0.293$.

Table E.8: Effect of log-probability recomputation and token-wise likelihood-ratio clipping.

| **Configuration** | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| NFPO | 0.133 | 0.074 | **0.081** | 0.496 | **0.745** | **0.262** | 0.360 | **0.307** |
| w/ recompute | 0.132 | 0.081 | 0.049 | 0.440 | 0.734 | 0.257 | 0.355 | 0.293 |
| w/o $\rho_t$ clip | **0.135** | **0.103** | 0.076 | **0.501** | 0.739 | 0.234 | **0.361** | **0.307** |

**Token-mask compatibility.** The TV-based mask achieves the highest average score, while NFPO can also be combined with KL, PPO, and IcePop-based masks. Table E.9 reports the full per-benchmark results.

Table E.9: Effect of the token-mask strategy used with NFPO.

| **Mask type** | AIME24 | AIME25 | AIME26 | AMC23 | MATH 500 | Minerva | Oly. | Avg. |
|---|---|---|---|---|---|---|---|---|
| TV | **0.133** | **0.074** | 0.081 | **0.496** | **0.745** | **0.262** | **0.360** | **0.307** |
| KL | 0.044 | 0.067 | **0.128** | 0.439 | 0.718 | 0.237 | 0.336 | 0.281 |
| PPO | 0.096 | 0.054 | 0.043 | 0.424 | 0.690 | 0.229 | 0.323 | 0.266 |
| IcePop | 0.087 | 0.036 | 0.033 | 0.417 | 0.686 | 0.229 | 0.313 | 0.258 |

### E.4 Computational and Memory Overhead

NFPO has peak VRAM usage and actor-update time comparable to DPPO, while reducing the total actor-update time by $20.2\%$ relative to GRPO. It requires no additional model forward or backward passes. Table E.10 reports the measurements using Qwen3-1.7B-Base with 16k-token sequences. DPPO and NFPO reuse the rollout log probabilities, whereas GRPO recomputes them in the training backend.

Table E.10: Peak VRAM and actor-update time using Qwen3-1.7B-Base with 16k-token sequences.

| **Cost** | **GRPO** | **DPPO** | **NFPO** ($N = 8$) |
|---|---|---|---|
| Peak VRAM per GPU | 45.96 GB | 45.97 GB | 45.97 GB |
| Actor forward/backward, including FSDP collectives | 36.34 s | 36.81 s | 36.86 s |
| Rollout log-probability recomputation | 9.84 s | 0 s | 0 s |
| Total actor-update time | 46.18 s | 36.81 s | 36.86 s |

## F Limitations and Broader Impacts

Our work primarily focuses on the RLVR setting with sparse, verifiable outcome rewards, which has recently emerged as a key paradigm for post-training of LLMs. While the proposed method is motivated by a general policy-improvement identity, extending the analysis and practical implementation to settings with dense or continuous reward signals may require additional empirical investigation.

This work studies RL algorithms for improving the reasoning capabilities of LLMs. More effective RL training methods may contribute positively to applications requiring reliable reasoning, including educational tools, coding assistants, and scientific support systems. At the same time, improving the reasoning ability of language models may also increase the capability of systems used for harmful purposes. Our work does not introduce new datasets or deployment systems, and our experiments are conducted using existing open-weight models and publicly available benchmarks. Nevertheless, as with other works in LLM training, the broader societal effects depend on how such models are deployed and governed in downstream applications.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the theoretical analysis, the proposed algorithm, and the empirical findings presented in Sections 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results state their assumptions explicitly, and complete proofs are provided in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the training framework, datasets, hyperparameters, and implementation details necessary to reproduce the experiments in Section 4 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available datasets and build on publicly available GitHub code. Code for our method is included in the supplementary material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides detailed descriptions of datasets, models, optimization settings, rollout configuration, hyperparameters, and evaluation protocols in Sections 4 and Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars or confidence intervals because multiple independent large-scale RLVR training runs are prohibitively expensive. We instead compare all methods under the same evaluation protocol and report results across multiple reasoning benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information on the computation resource in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Appendix F, although our work is primarily methodological and theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new large-scale models, datasets, or systems that require specific safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the datasets, models, and frameworks used in the experiments, including Qwen3 [44], MATH [12], DAPO-Math-17k [45], and verl [34].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code for our method in the supplementary material, together with the implementation details and hyperparameters needed to reproduce the experiments. We do not introduce new datasets or model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in the paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.