What are Key Factors for Updates in RL for LLM Reasoning?

Peidong Wang[1,3]* **Demi Wang**[2,3]* **Xufang Luo**[3]† **Jiahang Xu**[3]
Xiaocui Yang[1] **Shi Feng**[1]‡ **Yuqing Yang**[3] **Dongsheng Li**[3]

[1]Northeastern University
[2]Carnegie Mellon University
[3]Microsoft Research

Abstract

Reinforcement Learning from Verifiable Rewards (RLVR) has emerged as a promising framework for enhancing the reasoning ability of large language models. However, much of the existing work is guided by heuristic intuition, leading to divergent algorithmic choices, even contradictory ones that nevertheless report empirical gains. To better understand this phenomenon, we conduct a theoretical analysis of RLVR updates. Our study reveals that differences in off-policy degree, determined by the number of gradient steps per rollout, substantially affect the distribution of importance sampling ratios and their clipping behavior, thereby altering which tokens dominate the update. Building on this insight, we characterize gradient expectation as the central quantity governing update dynamics and analyze the roles of token probability, advantage, and importance sampling ratio. Motivated by these findings, we propose Adaptive Clip Policy Optimization (ACPO), which adjusts clipping boundaries across token groups according to the empirical variance of their importance sampling ratios. Experiments on 3B and 7B models across diverse reasoning benchmarks, spanning mathematical problem solving, tabular QA, and logic puzzles, demonstrate that ACPO outperforms strong baselines such as DAPO and CISPO. These results demonstrate that principled, analysis-inspired approaches yield more robust and effective RLVR methods. Code is available in: `https://github.com/Control-derek/ACPO`

1 Introduction

Reinforcement Learning from Verifiable Rewards (RLVR) aims to enhance the reasoning ability of large language models (LLMs) by optimizing against verifiable outcomes, such as mathematical correctness or logical validity [Shao et al., 2024, Guo et al., 2025a, Lambert et al., 2024, Jaech et al., 2024, Chen et al., 2025c, Yang et al., 2025a]. This setting provides a scalable way to directly reward structured reasoning, making RLVR a central approach for advancing reliability in complex problem-solving tasks.

Despite substantial progress, RLVR research remains largely heuristic-driven rather than systematic analysis. Accordingly, many algorithmic variants have been proposed [Yu et al., 2025, Yue et al., 2025, Su et al., 2025, Zheng et al., 2025a], yet their effectiveness is still not well understood. As illustrated in Figure 1(a), even seemingly contradictory token-level heuristics report similar gains: Wang et al. [2025b] prioritize high-entropy tokens and mask low-entropy ones, arguing that uncertain tokens

*Equal contribution. Work done during internship at Microsoft Research Asia.
†Corresponding author: `luoxufang@outlook.com`.
‡Corresponding author: `fengshi@cse.neu.edu.cn`.

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

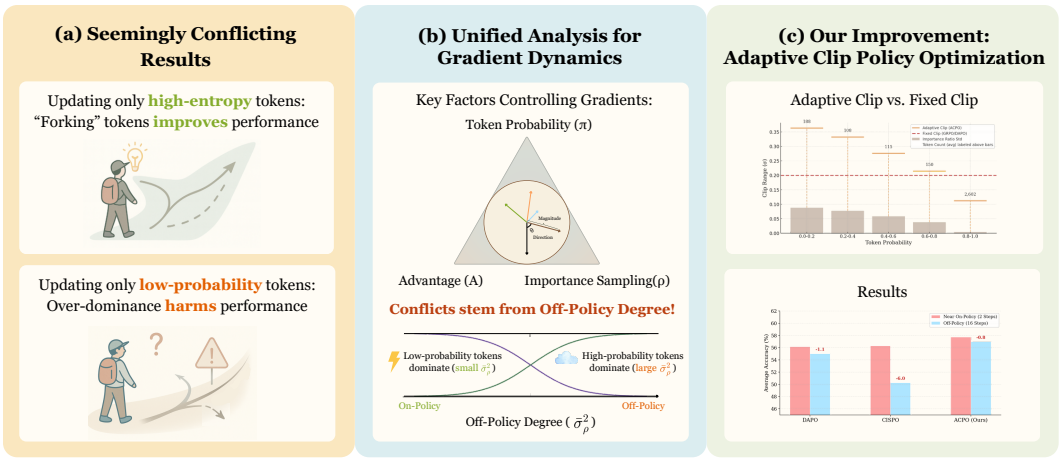

Figure 1: (a) Conflicting results: high-entropy updates help, while low-probability dominance hurts stability. (b) Our analysis attributes this to the clipped update, which shifts with off-policy degree via token probability, advantage, and IS. (c) ACPO adapts clipping per token group, improving over DAPO/CISPO across three reasoning tasks and off-policy regimes.

are pivotal for reasoning, whereas Yang et al. [2025b] emphasize high-probability tokens to prevent low-probability tokens from dominating gradient updates. Since token entropy is inversely correlated with token probability, these prescriptions are effectively opposite, yet both work in practice.

Motivated by the importance of token-level understanding in RLVR [Yang et al., 2025b, Wang et al., 2025b, Cui et al., 2025, Chen et al., 2025a, Wang et al., 2025a], we study the learning dynamics of clipped policy-gradient updates in RLVR. Our starting point is a simple but under-emphasized observation: seemingly contradictory token-selection heuristics are often evaluated under *different degrees of off-policy*. For instance, the high-entropy update strategy uses many more updates per rollout than the high-probability one, causing the optimized policy to drift further from the behavior policy that generated the rollouts. In practice, such off-policy regimes can be unavoidable because updates are often the bottleneck: one can cheaply generate many rollouts, but can only fit a limited number of update samples per step, making it natural to reuse each rollout across multiple updates.

Because GRPO [Shao et al., 2024] relies on a clipped importance-sampling (IS) ratio, off-policy degree affects both token-gradient magnitude and *which tokens survive clipping and contribute updates*. This implies that analyses based on per-token intuition can be misleading: the overall update is an aggregation over tokens, and clipping can systematically suppress different token groups, changing both the update magnitude and direction. Instead, one must examine the **expected effective gradient under clipping**. In Section 2, we quantify off-policy degree via the variance of the IS ratio and derive a closed-form characterization of the conditional gradient expectation. A key consequence is a **gradient dominance reversal**: in near on-policy regimes, low-probability tokens can dominate the effective update, while in sufficiently off-policy regimes their IS ratios become highly dispersed and are frequently clipped, causing higher-probability tokens to dominate. This reversal unifies the conflicting empirical findings in prior work.

Beyond resolving this contradiction, we ask more generally: what governs (i) which tokens contribute to the update, and (ii) how the update direction deviates from the on-policy ideal? In Section 3, we provide a component-wise analysis of RLVR updates along two dimensions: the *magnitude* of the effective update and its *directional bias* relative to the on-policy gradient. We show that token probability, advantage sign/magnitude, and IS-ratio clipping interact to shape both. These results highlight a fundamental mismatch in standard practice: a single global clipping window is not aligned with the highly heterogeneous IS-ratio dispersion across token groups and training regimes, and can also substantially rotate the optimization direction.

Motivated by these insights, we propose **Adaptive Clip Policy Optimization (ACPO)** in Section 4. ACPO replaces the fixed global clipping range with bin-wise thresholds that adapt to heterogeneous IS-ratio dispersion across token groups, reducing uneven clipping and mitigating clipping-induced gradient bias. Empirically, ACPO consistently outperforms strong baselines (e.g., DAPO [Yu et al.,

2025] and CISPO [Chen et al., 2025a]) across three reasoning tasks under both near on-policy and off-policy training regimes. In summary, our main contributions are as follows:

- Analysis of clipped RLVR: off-policy degree shifts gradient expectation via IS-ratio dispersion, causing gradient dominance reversal and reconciling prior heuristics.
- Joint effects of token probability, advantage, and IS-ratio clipping on update magnitude/direction, explaining why a single global clip can be mismatched.
- ACPO: variance-aware adaptive clipping with consistent gains over strong baselines across model scales, tasks, and both near on-policy and off-policy regimes.

2 Theoretical Analysis of Empirical Contradictions

Preliminary. In contrast to traditional actor-critic algorithms such as Proximal Policy Optimization (PPO) [Schulman et al., 2017], Group Relative Policy Optimization (GRPO) [Shao et al., 2024] removes the critic (of comparable size to the policy) and estimates advantages by standardizing rewards within a sampled group. For each prompt $q \sim P(Q)$, the policy π_θ samples G responses $\{o_i\}_{i=1}^G$, each scored by a rule-based reward $r_i = R(o_i)$. The relative advantage is $A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}$, where mean and std denote the sample mean and standard deviation, respectively. The objective is:

$$J_{GRPO}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^G\sim\pi_{\theta_{old}}(O|q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\{\min[\rho_{i,t}A_i, \text{clip}(\rho_{i,t}, 1-\epsilon_l, 1+\epsilon_h)A_i]\} \quad (1)$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$ is the importance sampling (IS) ratio, used to correct for off-policy updates. The clipping function $clip(\cdot, 1-\epsilon_l, 1+\epsilon_h)$ restricts this ratio to $[1-\epsilon_l, 1+\epsilon_h]$ for stable updates. GRPO typically uses symmetric clipping ($\epsilon_l = \epsilon_h$), while DAPO [Yu et al., 2025] adopts asymmetric clipping ($\epsilon_h > \epsilon_l$) to mitigate entropy collapse. Following recent practice [Yu et al., 2025, Su et al., 2025, Chen et al., 2025a], we omit the KL penalty.

2.1 Empirical Contradictions and the Role of Off-Policy Degree

Contradictory Empirical Results. The limitations of a heuristic-driven approach are highlighted by recent, conflicting findings. Wang et al. [2025b] report faster convergence and superior accuracy by selectively updating high-entropy tokens, as these tokens are pivotal for reasoning. In contrast, Yang et al. [2025b] argue against amplifying the influence of low-probability tokens, showing that exclusively updating them leads to slow convergence and suboptimal performance. Their work suggests that the large, high-variance gradients from these tokens can destabilize training. This creates a paradox, as high-entropy tokens are often those with low generation probabilities [Guo et al., 2025b, Zheng et al., 2025b]. That two strategies targeting overlapping token populations yield divergent outcomes suggests that a critical factor is being overlooked. This puzzling discrepancy motivates our investigation into the fundamental structure of the policy gradient.

Off-policy degree as the key. The key experimental difference between Yang et al. [2025b], Wang et al. [2025b] is the number of updates per rollout: 2 vs. 16. More updates push the current policy further away from the behavior policy that generated the rollouts, increasing the degree of off-policy. Because Eq. (1) uses a clipped IS ratio, off-policy degree affects not only the magnitude of token gradients but also which tokens survive clipping and contribute any update. Therefore, we analyze the **expected effective gradient under clipping**.

To quantify off-policy degree mechanistically, we track how dispersed the IS ratio is. Specifically, we model the conditional IS-ratio variance given the behavior-policy probability, $\sigma_\rho^2(\pi) = \text{Var}[\rho_{i,t} \mid \pi_{i,t} = \pi]$, and later aggregate it into a scalar proxy for the overall off-policy degree.

2.2 Detailed Analysis

Overview. We characterize which tokens dominate the effective GRPO update after clipping by studying a token-level update coefficient and its conditional expectation given the token's behavior-policy probability. Let $\pi_{i,t} \triangleq \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})$ for token position t in sampled sequence i. We show that increasing off-policy-ness (e.g., more updates per rollout) can reverse which probability region dominates the effective gradient, yielding a **Gradient Dominance Reversal** that reconciles Yang et al. [2025b], Wang et al. [2025b].

Gradient of GRPO. To connect our analysis to the GRPO objective, we start from its gradient and express it as a sum of token-level contributions:

$$\nabla_\theta J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} g_{i,t}(\theta),$$
$$g_{i,t}(\theta) \triangleq \rho_{i,t} A_i \cdot \mathbb{I}_{\text{clip}}(\rho_{i,t}, A_i) \cdot \nabla_\theta \log \pi_\theta(o_{i,t} \mid q, o_{i,<t}), \tag{2}$$

The clipping indicator is

$$\mathbb{I}_{\text{clip}}(\rho_{i,t}, A_i) = \begin{cases} 0 & A_i > 0,\ \rho_{i,t} > 1+\epsilon_h, \\ 0 & A_i < 0,\ \rho_{i,t} < 1-\epsilon_l, \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

From parameter gradients to an analyzable logit coefficient. Directly analyzing $g_{i,t}(\theta)$ at the parameter level is difficult because it entangles model Jacobians, so we instead analyze the initial update signal at the pre-softmax logit of the sampled token, where $\partial \log \pi_\theta(o_{i,t})/\partial z_{i,t,o_{i,t}} = 1 - \pi_\theta(o_{i,t})$. Under the standard small-step regime, the resulting token-wise effective logit-gradient coefficient can be approximated as

$$G_{i,t} \triangleq (1-\pi_{i,t}) \cdot \rho_{i,t} \cdot A_i \cdot \mathbb{I}_{\text{clip}}(\rho_{i,t}, A_i). \tag{4}$$

Here we use $\pi \in (0,1)$ to denote a specific value of random variable $\pi_{i,t}$ when conditioning. We study the conditional expectation $\mathbb{E}[G_{i,t} \mid \pi_{i,t} = \pi]$, abbreviated as $\mathbb{E}[G \mid \pi]$. The expectation is over the sampled prompts (and the induced randomness in A_i and $\rho_{i,t}$), conditioned on $\pi_{i,t} = \pi$.

In Eq. 4, a token contributes only if $\rho_{i,t}$ lies in the clipping interval $[1-\epsilon_l,\ 1+\epsilon_h]$ (with the boundary chosen based on the sign of A_i). Thus, analyzing $\mathbb{E}[G_{i,t} \mid \pi_{i,t} = \pi]$ reduces to characterizing how often $\rho_{i,t}$ falls inside this interval under the conditioning.

Conditioned on $\pi_{i,t} = \pi$, $\rho_{i,t}$ remains random because the current-policy probability $\pi_\theta(o_{i,t} \mid q, o_{i,<t})$ varies across prompts and evolves across updates. We denote this variability by the conditional distribution $p(\rho \mid \pi)$ and summarize its spread by the conditional variance

$$\sigma_\rho^2(\pi) \ \triangleq\ \mathrm{Var}[\rho_{i,t} \mid \pi_{i,t} = \pi]\,. \tag{5}$$

We treat $\sigma_\rho^2(\pi)$ as the local, token-group-specific measure of off-policy degree, and aggregate it across the rollout token-probability distribution $p(\pi)$ to obtain a single scalar measure of overall off-policy strength,

$$\bar{\sigma}_\rho^2 \ \triangleq\ \mathbb{E}_{\pi\sim p(\pi)}\left[\sigma_\rho^2(\pi)\right]. \tag{6}$$

As training becomes more off-policy, $p(\rho \mid \pi)$ becomes more dispersed and both $\sigma_\rho^2(\pi)$ and $\bar{\sigma}_\rho^2$ grow, which changes the expected amount of clipping and thus reshapes $\mathbb{E}[G \mid \pi]$. Throughout, we use $\sigma_\rho^2(\pi)$ as the local axis and $\bar{\sigma}_\rho^2$ as the global axis of off-policy degree.

The key is to characterize how $\sigma_\rho^2(\pi)$ depends on π under a small update step. Using a first-order approximation (Appendix C.1), we obtain:

Lemma 2.1 (Variance of Importance Sampling Ratio)**.** *Under a small logit perturbation Δy from one update step, for a token with probability π, the IS-ratio variance admits the approximation*

$$\sigma_\rho^2(\pi) = \kappa^2(1-\pi)^2 + O(\kappa^3), \tag{7}$$

where κ is an effective coefficient that summarizes the per-step logit-perturbation scale together with a vocabulary shape factor (see Appendix C.1). The higher-order term $O(\kappa^3)$ is negligible under the standard small-step assumption.

Eq. (7) shows that $\sigma_\rho^2(\pi)$ grows rapidly as π decreases, implying a much wider IS-ratio distribution for low-probability tokens and hence a higher likelihood of being clipped. Figure 2(a) provides an empirical visualization of this principle. The distribution of the IS ratio for low-probability tokens is shown to have a much larger spread and numerous outliers compared to that of high-probability tokens. We further fit $\mathrm{Var}(\rho)$ against $(1-\pi)$ and obtain an exponent of 1.90 ± 0.14, closely matching the theoretical value of 2 (Appendix Figure 4).

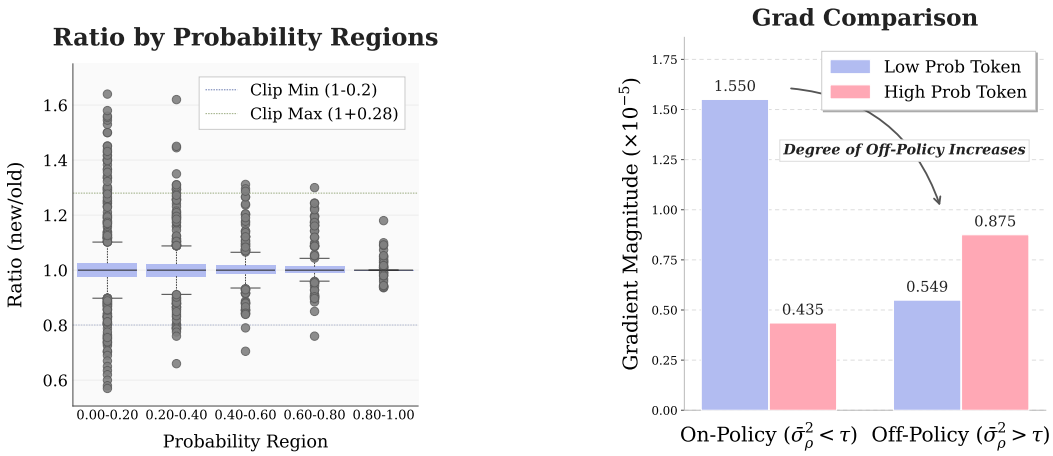

Figure 2: Empirical patterns of importance sampling ratio variance and gradient dominance. (a) Importance sampling ratios show higher variance for low-probability tokens. (b) As off-policy degree increases, gradient dominance shifts from low- to high-probability tokens.

Expectation of Gradient. Lemma 2.1 characterizes how the local off-policy degree $\sigma_\rho^2(\pi)$ depends on π. We next translate this into an analytical approximation for $\mathbb{E}[G \mid \pi]$ under three modeling choices: (i) a log-normal IS ratio, $\log \rho \mid \pi \sim \mathcal{N}\big(-\tau^2(\pi)/2,\ \tau^2(\pi)\big)$, with $\tau(\pi) := \sqrt{\mathrm{Var}(\log \rho \mid \pi)}$ and the monotone reparametrization $\tau^2(\pi) = \log(1 + \sigma_\rho^2(\pi))$; (ii) a probability-conditional linear coupling between advantage and the centered log-ratio with π-dependent slope $\mu(\pi) \geq 0$ (not a single global coupling; its empirical strength is measured in Appendix N.2); and (iii) the noiseless limit of this coupling for tractability. Full statements, finite-noise discussion, and a Gaussian re-derivation of the same qualitative conclusions are given in Appendix C.2 and Appendix C.2.9. Under these modeling choices, Appendix C.2 derives the following expression.

Proposition 2.2 (Conditional Gradient Expectation)**.** *Fix a token with behavior-policy probability* $\pi \in (0, 1)$ *and write* $\tau = \tau(\pi)$*. Define the standardized clip endpoints*

$$L(\pi) = \frac{\log(1-\epsilon_l) + \tau^2/2}{\tau}, \qquad U(\pi) = \frac{\log(1+\epsilon_h) + \tau^2/2}{\tau}, \qquad L'(\pi) = \min\big(L(\pi), 0\big). \tag{8}$$

The conditional gradient expectation admits the closed form

$$\mathbb{E}[G \mid \pi] \ = \ (1-\pi)\,\mu(\pi)\,H\big(\tau(\pi)\big), \tag{9}$$

where

$$H(\tau) \ = \ \tau\Big\{\phi\big(L'-\tau\big) - \phi\big(U-\tau\big) \ + \ \tau\big[\Phi(U-\tau) - \Phi(L'-\tau)\big]\Big\}. \tag{10}$$

The factor $(1-\pi)$ recovers the familiar on-policy logit-gradient front factor; $\mu(\pi)$ encodes the local coupling between advantage and the log IS-ratio; and $H(\tau)$ captures how the trust-region indicator selects which tokens contribute as the local off-policy degree grows. As $\tau \to 0$ (equivalently, $\sigma_\rho^2(\pi) \to 0$), $H(\tau) = \tau^2 + O(\tau^4) = \sigma_\rho^2(\pi) + O\big(\sigma_\rho^4(\pi)\big)$, so the gradient signal scales as $\mu(\pi)(1-\pi)\,\sigma_\rho^2(\pi)$ in the near-on-policy regime. As $\sigma_\rho^2(\pi)$ grows, the trust-region window $[L'(\pi), U(\pi)]$ moves into the tail of the tilted log-normal density, the indicator mass $\Phi(U-\tau) - \Phi(L'-\tau)$ shrinks, and $H(\tau)$ eventually decreases, suppressing the contribution of tokens with high local off-policy degree (typically low-probability tokens, by Lemma 2.1).

Two Different Gradient Dominances. Proposition 2.2 enables us to compare the expected update contributions from different token populations under clipping. Partition the token-probability space into a low-probability interval $I_L = [0, p_L]$ and a high-probability interval $I_H = [p_H, 1)$, with $0 < p_L < p_H < 1$, and let $\bar{G}_L(\bar{\sigma}_\rho^2)$ and $\bar{G}_H(\bar{\sigma}_\rho^2)$ denote the conditional averages of $\mathbb{E}[G \mid \pi]$ over $p(\pi)$ within I_L and I_H, respectively (full definitions and derivations in Appendix C.3). Both depend on the global off-policy degree $\bar{\sigma}_\rho^2$ through the local axis $\sigma_\rho^2(\pi)$ entering $\mathbb{E}[G \mid \pi]$. We now characterize how the dominant contribution shifts as $\bar{\sigma}_\rho^2$ grows.

Corollary 2.3 (Sufficient Condition for Gradient Dominance Reversal)**.** *Under the modeling choices in Proposition 2.2, define* $D(\bar{\sigma}_\rho^2) := \bar{G}_L(\bar{\sigma}_\rho^2) - \bar{G}_H(\bar{\sigma}_\rho^2)$*. Suppose* ***(C1)*** $\lim_{\bar{\sigma}_\rho^2 \to 0^+} D(\bar{\sigma}_\rho^2)/\bar{\sigma}_\rho^2 = C_0 > 0$*, where* C_0 *is a structural constant determined by* (μ, p, I_L, I_H) *(closed form in Appendix C.3); and* ***(C2)*** *there exists* $s_1 > 0$ *with* $D(s_1) < 0$*. Then there exists* $s_a \in (0, s_1)$ *with* $D(s_a) > 0$*, and by the intermediate value theorem at least one crossing point* $s^\star \in (s_a, s_1)$ *satisfies* $D(s^\star) = 0$*. We refer to such* $s^\star$ *as a* ***gradient dominance reversal threshold****: when* D *is sign-changing at* $s^\star$*, low-probability tokens dominate the conditional gradient signal in some left neighborhood of* $s^\star$ *and high-probability tokens dominate in some right neighborhood.*

Remark. (C1) and (C2) are concrete properties of D that are verifiable from the closed form once $p(\pi)$ and $\mu(\pi)$ are specified; we verify both for the representative setup of this paper in Appendix C.3.

Fig. 2(b) is consistent with this prediction (a controlled sweep over intermediate update counts is in Appendix N.1): tracking gradient norms of Qwen2.5-7B [Yang et al., 2024] trained on math reasoning with GRPO, low-probability tokens yield much larger gradients than high-probability tokens near on-policy (left), and as the off-policy degree increases (right) low-probability contributions collapse under clipping while high-probability tokens dominate.

Resolving the Empirical Contradiction. This dominance reversal unifies the conflicting observations: Yang et al. [2025b] operate at small $\bar{\sigma}_\rho^2$ where low-probability tokens dominate the gradient but inject high variance, so suppressing them stabilizes learning; Wang et al. [2025b] operate at larger $\bar{\sigma}_\rho^2$ where low-probability tokens are frequently clipped (Figure 2(a)), so the surviving high-entropy contributors are comparatively stable and informative.

3 Key Factors in LLM RL Updates

Section 2 shows that seemingly conflicting token-update heuristics can be reconciled once we analyze the effective update under clipped importance sampling: the off-policy degree changes the IS-ratio dispersion, which in turn changes which tokens survive clipping and dominate the expected gradient. To turn this insight into principled algorithm design, we further ask: beyond token probability, what factors control (i) which tokens contribute to the update, and (ii) how the update direction deviates from the on-policy ideal? This section provides a component-wise analysis of the GRPO update that will directly motivate adaptive clipping in Section 4.

3.1 Analytical Framework: magnitude vs. direction

We analyze GRPO updates from two perspectives.

Magnitude (who dominates the update). Following Section 2, we use the conditional gradient expectation $\mathbb{E}[G \mid \pi]$ (Proposition 2.2) to characterize *effective* token contributions after clipping, capturing both per-token gradient scaling and clipping probability.

Direction (how the optimization trajectory changes). Clipping is not merely a variance-reduction device: by zeroing out subsets of token-level updates, it can change the composition of the gradient and thus rotate the optimization direction even when the overall norm is similar. To quantify such directional changes in high dimensions, we adopt a layer-wise geometric measure: principal angles between the top-k right-singular subspaces of two layer-gradient matrices, reporting mean angles over the top-k directions (full procedure in Appendix D). Principal angles compare gradient subspaces and are invariant to sign flips, so we interpret them as changes in the dominant gradient subspace rather than in the signed update direction.

3.2 Key Factors

Eq. 4 shows that GRPO's *effective* gradient is jointly determined by (i) token-wise sensitivity and probability-dependent IS variability, (ii) the advantage magnitude and sign, and (iii) IS-ratio clipping that gates off-policy contributions. We analyze each factor through both gradient expectation $\mathbb{E}[G \mid \pi]$ (magnitude dominance) and subspace alignment (directional distortion).

Token properties: tail tokens are strong but fragile contributors. Low-probability tokens have large base sensitivity via $(1 - \pi)$, making them potentially dominant contributors in near on-policy regimes (Corollary 2.3). However, Lemma 2.1 shows that their IS-ratio variance grows rapidly as

π decreases, increasing their clipping probability as training becomes more off-policy. Empirically, under near on-policy training, gradients restricted to low-probability tokens exhibit disproportionately large norms and strong alignment with the full update (Appendix F.1), confirming that the learning signal is concentrated in the distribution tail, but that this signal is also the most susceptible to being filtered out by clipping. This clarifies why entropy-/probability-based heuristics can both help: they are implicitly trading off signal strength (tail tokens) against stability under clipping.

Advantage sign: negative-advantage samples steer the update direction. While advantage magnitude scales the update, its sign changes the qualitative effect on the policy. Positive-advantage updates ($A > 0$) reinforce sampled actions, whereas negative-advantage updates ($A < 0$) penalize specific actions and effectively redistribute probability mass across many alternatives. As a result, $A < 0$ samples can dominate the directional structure even when the norms are comparable. Consistent with this mechanism, we find that the negative-advantage gradient subspace aligns substantially better with the final update direction than the positive-advantage one (Appendix F.2). This highlights that RLVR optimization is often driven more by error correction than by simply reinforcing already-good trajectories, and it implies that any gating mechanism (e.g., clipping) that disproportionately suppresses $A < 0$ updates can significantly rotate the learning direction.

Importance sampling and clipping: stability comes with directional distortion. Importance sampling enables reuse of rollouts under policy drift, but the variance of ρ necessitates clipping for stable training. Crucially, clipping gates token updates through $\mathbb{I}_{\text{clip}}(\rho, A)$, which changes the effective sample set and can therefore distort the update direction. We empirically compare an on-policy gradient (aggregating samples into a single step) with the standard off-policy clipped gradient and observe a large directional divergence ($\approx 47^\circ$) despite similar magnitudes (Appendix F.3). This indicates that clipping does more than control variance: by down-weighting or discarding tokens with extreme IS ratios, it changes which tokens drive learning and can therefore shift the update direction away from the on-policy gradient, even when the gradient norm is similar.

3.3 Summary and design implication

Our analysis yields two algorithm design insights.

Uniform clipping is mismatched to heterogeneous IS-ratio dispersion. Since IS-ratio variance depends strongly on token probability (Lemma 2.1) and changes with the off-policy degree, a single global clipping window makes the probability of being clipped highly uneven across token groups, systematically over-clipping some groups (typically low-probability tokens) while letting others dominate the effective update.

Clipping affects not only variance but also the optimization trajectory. By changing which token updates survive, clipping can substantially rotate the gradient direction relative to the on-policy ideal, even when the overall gradient norm is preserved. This suggests that improving RLVR updates is not only about stabilizing magnitudes, but also about controlling which tokens steer the direction.

Implication: clipping should adapt to the local policy shift. Together, these results suggest that the clipping window should be calibrated per token group according to its off-policy degree, rather than being fixed globally. A natural, directly observable proxy is the within-group IS-ratio variance, $\sigma^2_\rho(\pi)$. This motivates the adaptive, variance-aware clipping strategy introduced in Section 4.

4 Adaptive Clip Policy Optimization

4.1 Motivation

GRPO reuses rollout trajectories via importance-sampling (IS) ratios and stabilizes learning with clipping. The gradient of the clipped objective (Eq. 2, cf. Eq. 4) implies that a token contributes to the update unless its IS ratio crosses the clipping boundary associated with the sign of its advantage, so learning depends critically on how dispersed IS ratios are across tokens. Lemma 2.1 provides an expression for the IS-ratio variance, implying substantially larger dispersion for low-probability tokens. With a single global clipping range, this leads to uneven clipping rates across token groups. Proposition 2.2 and Corollary 2.3 further show that such uneven clipping biases the expected gradient and can even change which token groups dominate the update, consistent with the update-direction distortions observed in Section 3.

Table 1: Countdown/HiTab results. Best validation reward, mean@4.

(a) Model trained on the Countdown dataset.

Method	Qwen2.5-3B-Instruct	
	N-OnP.	OffP.
Base Model	6.90	
AR-Lopti	65.82	57.41
High-Entropy	70.58	73.76
Low-Entropy	61.81	74.04
DAPO	73.27	74.38
CISPO	74.25	55.12
ACPO	**75.74**	**76.27**

(b) Model trained on the HiTab dataset.

Method	Qwen2.5-3B		Qwen2.5-7B	
	N-OnP.	OffP.	N-OnP.	OffP.
Base Model	17.50		31.25	
AR-Lopti	45.50	43.33	68.92	66.58
High-Entropy	45.00	43.50	64.33	61.83
Low-Entropy	43.17	43.00	67.00	65.08
DAPO	44.33	44.00	68.42	65.75
CISPO	43.92	44.83	69.17	65.67
ACPO	**45.50**	**46.42**	**69.83**	**66.58**

Table 2: Qwen2.5-7B trained on the ORZ-57K. AIME24/25: avg@32; others: avg@8.

Method	Minerva		Math500		AMC		AIME24		AIME25		Olympiad		Avg.	
	N-OnP.	OffP.	N-OnP.	OffP.	N-OnP.	OffP.	N-OnP.	OffP.	N-OnP.	OffP.	N-OnP.	OffP.	N-OnP.	OffP.
Base Model	13.32		41.48		26.20		0.83		0.62		18.02		16.75	
AR-Lopti	29.14	25.14	73.75	63.92	44.88	34.94	15.21	8.75	3.96	2.08	33.60	24.00	33.42	26.47
High-Entropy	29.08	29.27	76.88	75.22	50.75	46.39	14.37	14.06	7.81	**10.62**	33.54	34.54	35.41	35.02
Low-Entropy	28.72	25.05	75.85	72.47	48.95	40.81	16.04	11.15	9.48	4.58	34.69	26.61	35.62	30.11
DAPO	29.92	30.47	78.05	76.72	51.20	46.84	**19.69**	**18.02**	11.88	6.56	41.44	37.20	38.70	35.97
CISPO	30.51	29.55	78.57	76.30	49.25	46.08	18.65	13.44	10.00	8.33	40.61	39.10	37.93	35.47
ACPO (Ours)	**31.02**	**31.66**	**79.78**	**80.48**	**54.37**	**51.20**	16.77	16.35	**14.37**	10.10	**42.83**	**42.17**	**39.86**	**38.66**

4.2 Adaptive Clip Policy Optimization (ACPO)

ACPO replaces the single global clipping range in Eq. 1 with a bin-specific range determined by the behavior-token probability. Given an update mini-batch (pooling all tokens in the update batch), we assign each token (i,t) to one of B equal-width bins on $[0,1]$ via

$$\mathcal{B}(i,t) = \min\{B, \lfloor B \cdot \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t}) \rfloor + 1\}. \tag{11}$$

For each bin b, we compute the within-bin IS-ratio dispersion and set the clipping range as

$$\sigma_b = \text{Std}\big(\{\rho_{i,t} : \mathcal{B}(i,t) = b\}\big), \qquad \epsilon_b = \text{clip}\big(\epsilon_{\text{base}} + \alpha\,\sigma_b,\ \epsilon_{\min},\ \epsilon_{\max}\big), \tag{12}$$

with $\epsilon_{\min} = 0$ and $\epsilon_{\max} = 3$ in all experiments (Appendix I). We apply token-wise clipping by using $\epsilon_{\mathcal{B}(i,t)}$ in Eq. 1. The objective is

$$J_{\text{ACPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big[\rho_{i,t}A_i,\ \text{clip}\big(\rho_{i,t}, 1-\epsilon_{\mathcal{B}(i,t)}, 1+\epsilon_{\mathcal{B}(i,t)}\big)A_i\Big], \tag{13}$$

See Appendix G (Algorithm 1) for details.

5 Experiments

5.1 Experimental Setup

Tasks, models, and training regimes. We evaluate ACPO on three reasoning tasks spanning different domains: ORZ-57K [Hu et al., 2025] for mathematical problem solving, HiTab [Cheng et al., 2022] for tabular question answering, and Countdown [Jackson, 2025] for arithmetic-based puzzles. Experiments use Qwen2.5-3B, Qwen2.5-3B-Instruct, and Qwen2.5-7B [Yang et al., 2024]. Following Yang et al. [2025b], Wang et al. [2025b], each method is run under a near on-policy regime (N-OnP., 2 updates/rollout) and an off-policy regime (OffP., 16 updates/rollout) to test robustness across policy-divergence levels.

Baselines, evaluation, and hyperparameters. We compare against: (1) **DAPO** [Yu et al., 2025], asymmetric clipping with $\epsilon_h > \epsilon_l$; (2) **CISPO** [Chen et al., 2025a], which clips only IS ratios rather than gating token updates; (3) **High/Low-Entropy** [Wang et al., 2025b], updating only top-20%/bottom-80% entropy tokens; and (4) **AR-Lopti** [Yang et al., 2025b], which scales token-wise

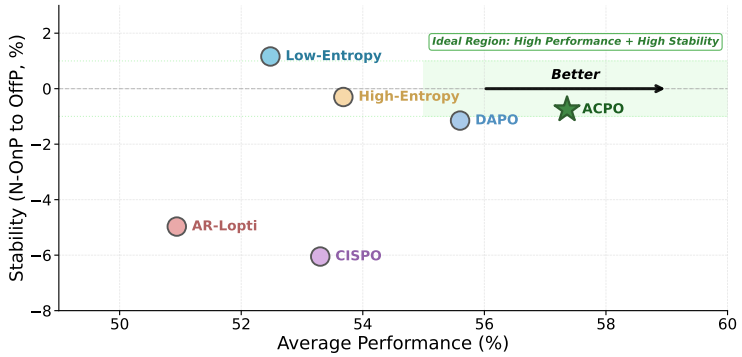

Figure 3: Performance vs. cross-regime stability. Each point averages four task–model settings: ORZ-57K with Qwen2.5-7B (mean of six benchmarks), Countdown with Qwen2.5-3B-Instruct, and HiTab with Qwen2.5-3B and 7B. x: mean of N-OnP and OffP performance; y: OffP minus N-OnP.

contributions to reduce low-probability dominance. For ORZ models we evaluate on six math benchmarks (Minerva [Lewkowycz et al., 2022], Math500 [Hendrycks et al., 2021], AMC2023, AIME24/25, OlympiadBench [He et al., 2024]), reporting avg@32 for AIME and avg@8 for the rest; HiTab and Countdown report the best validation reward (mean@4). Statistical significance uses one-tailed Welch's t-test ($p < 0.05$). ACPO uses $B = 5$ probability bins, $\alpha = 3$, and $\epsilon_{\text{base}} = 0.2$, with all methods sharing the same base configuration; full hyperparameters and per-method configs are in Appendix I, sensitivity in Appendix J.1, and full p-values in Appendix M.

5.2 Main Results

Consistent improvements across settings. Tables 1–2 present results across three reasoning tasks (ORZ, HiTab, Countdown), multiple model scales (3B, 7B), and two training regimes (near on-policy and off-policy). ACPO achieves the best average rank of 1.22 across all 18 task-regime configurations, with 15/18 first-place (including 2 ties with AR-Lopti) and 2/18 second-place finishes, compared to DAPO (avg rank 2.56, 2/18 wins) and CISPO (avg rank 3.11, 0/18 wins). One-tailed Welch's t-tests show that ACPO's gains are statistically significant ($p < 0.05$) on most benchmarks (Appendix M).

Cross-regime stability. ACPO remains stable when moving from near on-policy to off-policy, with only a 0.8-point performance drop (Figure 3); among baselines, CISPO and AR-Lopti suffer substantial degradation (-6.0 and -5.0 points), DAPO shows a moderate drop (-1.2 points), and High/Low-Entropy appear stable but at significantly lower overall performance. A plausible explanation for CISPO's instability is that it clips IS-ratio values without gating gradient updates, and for AR-Lopti that suppressing low-probability tokens becomes less helpful once gradient dominance reverses (Corollary 2.3). These results are consistent with the view that effective RLVR benefits from adapting *which tokens survive clipping* as off-policy degree changes, rather than relying on fixed heuristics.

Analysis of clipping and gradient dynamics. We further analyze ACPO's clipping behavior and gradient dynamics across probability bins (Appendix K, Figure 7): ACPO yields more uniform clip rates and substantially more stable gradients than DAPO, explaining its cross-regime stability. Appendices J.2–L further confirm that gains stem from adaptive calibration rather than simply using different threshold values, with negligible computational overhead.

6 Related Works

Reinforcement Learning for Large Language Models. RL for LLMs has shifted from preference alignment via RLHF [Ouyang et al., 2022] to directly optimizing reasoning abilities with Reinforcement Learning from Verifiable Rewards (RLVR) [Lambert et al., 2024]. This paradigm uses objective feedback signals, such as mathematical correctness, to train models. Pioneered by early work like OpenAI's o1 [Jaech et al., 2024] and advanced by algorithms such as GRPO [Shao et al., 2024], RLVR has become a central methodology for developing state-of-the-art reasoning agents, demonstrating that complex skills can emerge from outcome-based optimization.

Token-Level Update Strategies. RLVR's success depends heavily on the token-level update mechanism. Diverse heuristics have been proposed, including entropy-based modulation [Wang et al., 2025b, Cui et al., 2025], gradient clipping [Chen et al., 2025a, Zheng et al., 2025a], and advantage design, yet often yield conflicting conclusions. For instance, a key debate is whether to prioritize high-entropy tokens for exploration [Wang et al., 2025b] or suppress them to reduce gradient variance [Yang et al., 2025b]. Such contradictions reveal the lack of a unified understanding of RLVR update dynamics, motivating our systematic component-wise analysis.

7 Conclusion

We present a token-level analysis of clipped RLVR updates and identify off-policy degree as the key factor that reconciles prior conflicting heuristics. As off-policy increases, IS-ratio dispersion and clipping reshuffle which tokens dominate the expected effective gradient, revealing the mismatch of uniform clipping. This motivates ACPO, which uses variance-aware adaptive clipping thresholds.

Experiments across model scales (3B/7B), diverse reasoning tasks, and training regimes demonstrate that ACPO achieves consistent improvements and robust cross-regime stability, supporting our central thesis that understanding token-level gradient dynamics and adapting which tokens survive clipping accordingly, rather than simply relying on fixed heuristics, is key to effective RL for LLM reasoning.

Acknowledgments and Disclosure of Funding

This paper is supported by the National Natural Science Foundation of China (Project Nos. 62272092, 62676083 and 62606078), the General Project of the China Postdoctoral Science Foundation (Project No. 2026M791611), the Liaoning Provincial Doctoral Research Start-up Fund Project (Project No. 2026-BS-0112), and the Fundamental Research Funds for the Central Universities (Project No. N25XQD004). This work was done while Peidong Wang and Demi Wang were interns at Microsoft Research Asia.

References

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Rémi Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards. *arXiv preprint arXiv:2506.20520*, 2025.

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*, 2025a.

Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization. *arXiv preprint arXiv:2505.12346*, 2025b.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *arXiv preprint arXiv:2505.16400*, 2025c.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. HiTab: A hierarchical table dataset for question answering and natural language generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.78. URL `https://aclanthology.org/2022.acl-long.78/`.

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

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.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025.

Zitian Gao, Lynx Chen, Haoming Luo, Joey Zhou, and Bryan Dai. One-shot entropy minimization. *arXiv preprint arXiv:2505.20282*, 2025.

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

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. *arXiv preprint arXiv:2505.23564*, 2025b.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL `https://arxiv.org/abs/2402.14008`.

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. *arXiv preprint arXiv:2103.03874*, 2021.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL `https://arxiv.org/abs/2503.24290`.

Alex Jackson. Countdown numbers game dataset, 2025. Released under the MIT License.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. URL `https://arxiv.org/abs/2411.15124`.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857, 2022.

Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. Disco: Reinforcing large reasoning models with discriminative constrained optimization. *arXiv preprint arXiv:2505.12366*, 2025.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*, 2025.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

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.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy reinforce: Stable and efficient reinforcement learning for llms. *arXiv preprint arXiv:2503.14286*, 2025.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Zhenpeng Su, Leiyu Pan, Xue Bai, Dening Liu, Guanting Dong, Jiaming Huang, Wenping Hu, and Guorui Zhou. Klear-reasoner: Advancing reasoning capability via gradient-preserving clipping policy optimization. *arXiv preprint arXiv:2508.07629*, 2025.

Qwen Team. Qwq-32b: Embracing the Power of Reinforcement Learning. `https://qwenlm.github.io/blog/qwq-32b/`, March 2025. Accessed: 2025-09-14.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr. *arXiv preprint arXiv:2507.15778*, 2025a.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025b.

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.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in rl for llms. *arXiv preprint arXiv:2505.12929*, 2025b.

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.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, et al. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017*, 2025b.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*, 2025.

A Reproducibility

To ensure the reproducibility of our findings, we provide comprehensive details of our training process, including all hyperparameters and experimental configurations, in Appendix I.

B Ethics Statement

The primary objective of this work is to enhance the systematic understanding and reliability of reasoning in large language models. We view this as a crucial contribution to the development of safer and more predictable AI. All experiments were conducted on publicly available, standard academic datasets and utilized open-source base models. These resources do not contain any personally identifiable information or sensitive user data. While the base models may inherit societal biases from their pre-training data, our research focuses on optimizing for objective, verifiable tasks such as mathematical and logical correctness. This problem formulation is less susceptible to subjective biases, though we acknowledge that addressing underlying model bias remains a critical and ongoing challenge for the community. We also recognize that advancements in AI capabilities carry a potential for dual use, and we encourage the responsible development and deployment of these technologies. Our work is intended for a research audience to foster a deeper, more principled understanding of AI systems.

C Detailed Theoretical Derivations

C.1 Derivation of Lemma 2.1

This appendix derives the relationship between the variance of the importance sampling ratio, σ_ρ^2, and the token probability π_k in Lemma 2.1.

C.1.1 A small-step approximation in logit space

Let π_k be produced by a softmax with temperature T over logits $y \in \mathbb{R}^V$: $\pi_k = \frac{\exp(y_k/T)}{\sum_j \exp(y_j/T)}$. A single update step perturbs the entire logit vector by $\Delta y \in \mathbb{R}^V$, so $\Delta\pi_k$ is driven by changes to *all* logits, not just y_k.

The relevant softmax partial derivatives are

$$\frac{\partial \pi_k}{\partial y_k} = \frac{\pi_k(1-\pi_k)}{T}, \qquad \frac{\partial \pi_k}{\partial y_j} = -\frac{\pi_k\,\pi_j}{T} \quad (j \neq k). \tag{14}$$

Combining the contributions of all logits, the first-order expansion is

$$\Delta\pi_k \;\approx\; \frac{\pi_k}{T}\Big[(1-\pi_k)\,\Delta y_k - \sum_{j\neq k} \pi_j\,\Delta y_j\Big]. \tag{15}$$

The IS ratio for token k is $\rho_k = \pi_k'/\pi_k = 1 + \Delta\pi_k/\pi_k$, so

$$\rho_k \;\approx\; 1 + \frac{1}{T}\Big[(1-\pi_k)\,\Delta y_k - \sum_{j\neq k} \pi_j\,\Delta y_j\Big]. \tag{16}$$

C.1.2 Variance of the Importance Sampling Ratio

We now compute the variance of ρ_k from Eq. (16). The IS ratio ρ_k is a random variable whose randomness comes from two sources: (i) the per-step logit perturbation Δy, and (ii) the policy state itself, i.e., the full vocabulary distribution $\boldsymbol{\pi} = (\pi_k, \boldsymbol{\pi}_{\neq k})$ that varies across prompts and positions. We will first take variance with respect to Δy at a fixed $\boldsymbol{\pi}$, then marginalize over $\boldsymbol{\pi}_{\neq k}$ to obtain a single scalar function of π.

Assumption (per-step logit perturbation). We treat the per-step logit perturbation Δy as a random vector with zero mean, common per-coordinate variance σ_y^2, and uncorrelated coordinates:

$$\mathbb{E}[\Delta y_j] = 0, \quad \mathrm{Var}(\Delta y_j) = \sigma_y^2, \quad \mathrm{Cov}(\Delta y_j, \Delta y_{j'}) = 0 \ \ (j \neq j'). \tag{17}$$

The scalar σ_y^2 summarizes the magnitude of one update step in logit space and serves as a single per-step drift scale.

Step 1: conditional variance given the policy state. Conditioning on the policy state $\boldsymbol{\pi}$, the variance of Eq. (16) over Δy is

$$\sigma_\rho^2(\boldsymbol{\pi}) \triangleq \mathrm{Var}(\rho_k \mid \boldsymbol{\pi}) = \frac{1}{T^2}\Big[(1-\pi_k)^2 \,\mathrm{Var}(\Delta y_k) + \sum_{j \neq k} \pi_j^2 \,\mathrm{Var}(\Delta y_j)\Big] + R \tag{18}$$

$$= \frac{\sigma_y^2}{T^2}\Big[(1-\pi_k)^2 + \sum_{j \neq k} \pi_j^2 \Big] + R, \tag{19}$$

where R is the contribution of higher-order Taylor terms in Δy (discussed below).

Define the *shape factor* of the residual distribution over non-sampled tokens by

$$S(\boldsymbol{\pi}_{\neq k}) \ \triangleq \ \sum_{j \neq k} q_j^2, \qquad q_j \ \triangleq \ \frac{\pi_j}{1-\pi_k} \quad (j \neq k), \tag{20}$$

so $\{q_j\}_{j \neq k}$ is a normalized distribution and $S \in [\,1/(V-1),\, 1\,]$. Substituting $\sum_{j \neq k} \pi_j^2 = (1-\pi_k)^2\, S$ gives the factorized form

$$\sigma_\rho^2(\boldsymbol{\pi}) \ = \ \frac{\sigma_y^2}{T^2}\,(1-\pi_k)^2\,\big[1 + S(\boldsymbol{\pi}_{\neq k})\big] + R, \tag{21}$$

which separates a clean π_k-scaling, $(1-\pi_k)^2$, from a bounded distribution-shape factor $1+S \in [1,2]$.

Step 2: order of the remainder R. Writing the full Taylor expansion of ρ_k in Δy as $\rho_k = 1 + L(\Delta y) + Q(\Delta y) + \cdots$ with L linear and Q quadratic, we have

$$\mathrm{Var}(\rho_k \mid \boldsymbol{\pi}) = \mathrm{Var}(L) + 2\,\mathrm{Cov}(L, Q) + \mathrm{Var}(Q) + \cdots \tag{22}$$

The leading term $\mathrm{Var}(L)$ is $O(\sigma_y^2)$ and gives Eq. (21). The cross term $\mathrm{Cov}(L, Q)$ is controlled by the third moment of Δy and is $O(\sigma_y^3)$; $\mathrm{Var}(Q)$ is $O(\sigma_y^4)$. Hence $R = O(\sigma_y^3)$ in general, and degenerates to $O(\sigma_y^4)$ when Δy is symmetric.

Step 3: marginalize over $\boldsymbol{\pi}_{\neq k}$ to obtain $\sigma_\rho^2(\pi)$. The shape factor $S(\boldsymbol{\pi}_{\neq k})$ depends on the full vocabulary distribution, not just on π_k. To obtain the single-variable summary $\sigma_\rho^2(\pi)$ used in the main text, we average Eq. (21) over the conditional distribution of $\boldsymbol{\pi}_{\neq k}$ given $\pi_k = \pi$:

$$\sigma_\rho^2(\pi) \ \triangleq \ \mathbb{E}\big[\,\sigma_\rho^2(\boldsymbol{\pi}) \,\big|\, \pi_k = \pi\,\big] \ = \ \frac{\sigma_y^2}{T^2}\,(1-\pi)^2\,\big(1 + \bar{S}(\pi)\big) + \mathbb{E}[R \mid \pi_k = \pi], \tag{23}$$

where $\bar{S}(\pi) \triangleq \mathbb{E}[S(\boldsymbol{\pi}_{\neq k}) \mid \pi_k = \pi] \in [\,1/(V-1),\, 1\,]$.

Because $1 + \bar{S}(\pi)$ is a bounded factor in $[1, 2]$ and varies mildly compared with $(1-\pi)^2$, we absorb its (overall) magnitude into an *effective* per-step logit-perturbation scale,

$$\kappa^2 \ \triangleq \ \frac{\sigma_y^2}{T^2}\big(1 + \bar{S}\big), \qquad \bar{S} \triangleq \mathbb{E}_\pi[\bar{S}(\pi)], \tag{24}$$

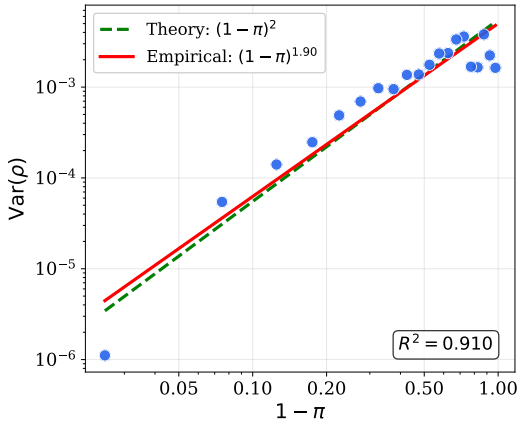

Figure 4: Empirical validation of Lemma 2.1 via log-log regression. Green: theoretical scaling $\mathrm{Var}(\rho) \propto (1-\pi)^2$. Red: fitted curve $\mathrm{Var}(\rho) = 5.01 \times 10^{-3}(1-\pi)^{1.90}$ (slope 1.90 ± 0.14, $R^2 = 0.91$).

which yields an exact two-sided bound up to a factor of two,

$$\tfrac{1}{2}\,\kappa^2(1-\pi)^2 \;\le\; \frac{\sigma_y^2}{T^2}(1-\pi)^2 \;\le\; \sigma_\rho^2(\pi) - \mathbb{E}[R \mid \pi_k = \pi] \;\le\; \frac{2\sigma_y^2}{T^2}(1-\pi)^2 \;\le\; 2\kappa^2(1-\pi)^2. \quad (25)$$

With this convention,

$$\sigma_\rho^2(\pi) \;=\; \kappa^2(1-\pi)^2 \;+\; O(\kappa^3), \quad (26)$$

which matches the form stated in Lemma 2.1. The $O(\kappa^3)$ remainder collects (i) the higher-order Taylor remainder $\mathbb{E}[R \mid \pi_k = \pi]$ from Step 2, and (ii) the residual variation of $1 + \bar{S}(\pi)$ around its average $1 + \bar{S}$. We treat the shape factor as constant in π (absorbed into κ); its variation with π is an additional modeling approximation, and both terms are neglected under the standard small-step assumption.

The leading term shows that σ_ρ^2 decreases (approximately) monotonically with π on $(0, 1)$: low-probability tokens have substantially wider IS-ratio distributions and are thus more susceptible to clipping.

C.1.3 Empirical validation of the variance scaling

To validate the prediction $\mathrm{Var}(\rho) \propto (1-\pi)^2$, we perform a log-log regression on binned tokens. As shown in Figure 4, the fitted exponent is 1.90 ± 0.14 with $R^2 = 0.91$ (20 bins; 54,173 tokens), which closely matches the theoretical exponent of 2. The green curve indicates the theoretical scaling, while the red curve shows the empirical fit:

$$\mathrm{Var}(\rho) = 5.01 \times 10^{-3} \cdot (1-\pi)^{1.90}. \quad (27)$$

C.2 Derivation of Proposition 2.2

This appendix derives the closed form for $\mathbb{E}[G \mid \pi]$ stated in Proposition 2.2. We work under the three modeling assumptions stated in Section 2.2, restated here for convenience:

(M1) *Log-normal IS-ratio.* $\log \rho \mid \pi \sim \mathcal{N}\big(-\tau^2(\pi)/2,\, \tau^2(\pi)\big)$, with $\tau^2(\pi) = \log\big(1 + \sigma_\rho^2(\pi)\big)$. The log-normal choice is for closed-form convenience; Appendix C.2.9 reproduces the qualitative conclusions (small-$\bar{\sigma}_\rho^2$ scaling, dominance reversal) under a Gaussian IS-ratio model.

(M2) *Linear-coupling ansatz.* $A = \mu(\pi)\, u + \eta$, where $u := \log \rho + \tau^2(\pi)/2$ is the centered log-ratio, $\mu(\pi) \ge 0$, and $\mathbb{E}[\eta \mid u, \pi] = 0$.

(M3) *Perfect-coupling idealization.* The closed form below is computed under $\eta \equiv 0$ (the finite-η effect is discussed at the end of this section).

C.2.1 Setup and the trust-region indicator

The token-level effective coefficient (Eq. 4) is

$$G \;=\; (1-\pi)\,\rho\,A\,\mathbb{I}_{\text{trust}}(\rho,A), \qquad \mathbb{I}_{\text{trust}}(\rho,A)=\begin{cases}1, & A>0,\ \rho\le 1+\epsilon_h,\\ 1, & A<0,\ \rho\ge 1-\epsilon_l,\\ 0, & \text{otherwise.}\end{cases} \tag{28}$$

Note that the indicator depends on A through $\operatorname{sign}(A)$: positive advantages are clipped from above, negative advantages from below. Under (M3), $\eta \equiv 0$ implies $A = \mu(\pi)\,u$ with $\mu(\pi) \ge 0$, so $\operatorname{sign}(A) = \operatorname{sign}(u)$ wherever $\mu(\pi) > 0$.[4] Therefore, the indicator collapses to a condition on the centered log-ratio u alone:

$$\mathbb{I}_{\text{trust}}\Big|_{\eta\equiv 0} \;=\; \mathbf{1}\{u>0,\ \log\rho\le\log(1+\epsilon_h)\} \;+\; \mathbf{1}\{u<0,\ \log\rho\ge\log(1-\epsilon_l)\}. \tag{29}$$

C.2.2 Standardization

Under (M1), the centered log-ratio $u = \log\rho + \tau^2/2$ satisfies $u \mid \pi \sim \mathcal{N}(0,\tau^2)$. Define the standardized variable

$$Z \;:=\; u/\tau \;=\; \big(\log\rho+\tau^2/2\big)/\tau \;\sim\; \mathcal{N}(0,1). \tag{30}$$

The two log-ratio thresholds $\log(1\pm\epsilon)$ map to standardized endpoints

$$L(\pi) \;=\; \frac{\log(1-\epsilon_l)+\tau^2/2}{\tau}, \qquad U(\pi) \;=\; \frac{\log(1+\epsilon_h)+\tau^2/2}{\tau}. \tag{31}$$

For typical clip parameters and small τ, $L(\pi) < 0 < U(\pi)$. We do not assume this in general; we use $L'(\pi) := \min(L(\pi), 0)$ to keep the lower piece of the integration domain inside $\{u < 0\}$.

Substituting $\rho = \exp(\tau Z - \tau^2/2)$ and $A = \mu(\pi)\,\tau Z$ into Eq. (29), the integration region in Z becomes

$$\Omega(\pi) \;=\; \{Z : 0 < Z \le U(\pi)\} \;\cup\; \{Z : L'(\pi) \le Z < 0\}. \tag{32}$$

(When $L \ge 0$, the lower piece $\{L \le Z < 0\}$ is empty; using $L' = \min(L, 0)$ keeps the closed form uniform across the regimes $L < 0$ and $L \ge 0$.)

C.2.3 Computing $\mathbb{E}[G \mid \pi]$

Under (M3), Eq. (29) and the substitutions above give

$$\begin{aligned}\mathbb{E}[G\mid\pi] &= (1-\pi)\,\mathbb{E}\Big[\rho\cdot A\cdot\mathbb{I}_{\text{trust}}\,\Big|\,\pi\Big]\\ &= (1-\pi)\,\mu(\pi)\,\mathbb{E}\Big[\,e^{\tau Z-\tau^2/2}\cdot\tau Z\cdot\mathbf{1}_{Z\in\Omega(\pi)}\,\Big]\\ &= (1-\pi)\,\mu(\pi)\,\tau\,\underbrace{\int_{\Omega(\pi)} z\,e^{\tau z-\tau^2/2}\,\phi(z)\,dz}_{=:\,J(\tau,\pi)}.\end{aligned} \tag{33}$$

We compute $J(\tau,\pi)$ using the completion-of-square identity for the standard normal density:

$$e^{\tau z}\,\phi(z) \;=\; \frac{1}{\sqrt{2\pi}}\exp\Big(\tau z-\tfrac{z^2}{2}\Big) \;=\; \frac{1}{\sqrt{2\pi}}\exp\Big(\tfrac{\tau^2}{2}-\tfrac{(z-\tau)^2}{2}\Big) \;=\; e^{\tau^2/2}\,\phi(z-\tau). \tag{34}$$

Therefore $e^{\tau z-\tau^2/2}\phi(z) = \phi(z-\tau)$, and

$$J(\tau,\pi) \;=\; \int_{\Omega(\pi)} z\,\phi(z-\tau)\,dz. \tag{35}$$

[4] If $\mu(\pi) = 0$ on a measure-zero set, G vanishes there and the closed form is unaffected.

Using the elementary identity $\int z\,\phi(z-\tau)\,dz = -\phi(z-\tau) + \tau\,\Phi(z-\tau) + C$, we evaluate the two pieces of $\Omega(\pi) = [L'(\pi), 0) \cup (0, U(\pi)]$:

$$\int_{L'(\pi)}^{0} z\,\phi(z-\tau)\,dz = \big[-\phi(z-\tau) + \tau\Phi(z-\tau)\big]_{L'(\pi)}^{0}$$
$$= -\phi(-\tau) + \tau\Phi(-\tau) \;+\; \phi(L'-\tau) - \tau\Phi(L'-\tau), \tag{36}$$

$$\int_{0}^{U(\pi)} z\,\phi(z-\tau)\,dz = -\phi(U-\tau) + \tau\Phi(U-\tau) \;+\; \phi(-\tau) - \tau\Phi(-\tau). \tag{37}$$

Adding the two pieces, the boundary contributions at $z = 0$ cancel exactly, yielding

$$J(\tau,\pi) \;=\; \phi(L'(\pi)-\tau) - \phi(U(\pi)-\tau) \;+\; \tau\big[\Phi(U(\pi)-\tau) - \Phi(L'(\pi)-\tau)\big]. \tag{38}$$

C.2.4 The closed form $H(\tau)$

Combining Eq. (33) and Eq. (38), we obtain

$$\mathbb{E}[G \mid \pi] \;=\; (1-\pi)\,\mu(\pi)\,H\big(\tau(\pi)\big), \quad H(\tau) \;=\; \tau\Big\{\phi(L'-\tau) - \phi(U-\tau) + \tau\big[\Phi(U-\tau) - \Phi(L'-\tau)\big]\Big\}, \tag{39}$$

which matches Proposition 2.2. The three factors admit the following interpretation: $(1-\pi)$ is the on-policy logit-gradient front factor; $\mu(\pi)$ is the local advantage–log-ratio coupling under (M2); $H(\tau)$ aggregates the trust-region indicator's effect on the tilted normal density (whose mean is shifted from 0 to τ by the IS factor) over the standardized window $[L', U]$.

C.2.5 Small-τ Taylor expansion

For small τ, the standardized endpoints behave as

$$U(\pi) \;=\; \frac{\log(1+\epsilon_h)}{\tau} + \frac{\tau}{2} \;\to\; +\infty, \qquad L(\pi) \;=\; \frac{\log(1-\epsilon_l)}{\tau} + \frac{\tau}{2} \;\to\; -\infty, \tag{40}$$

so $\Phi(U-\tau) \to 1$, $\Phi(L'-\tau) \to 0$, and $\phi(U-\tau), \phi(L'-\tau) \to 0$ super-polynomially. Hence

$$H(\tau) \;=\; \tau \cdot \tau \cdot \big(1-0\big) + o(\tau^2) \;=\; \tau^2 + O(\tau^4). \tag{41}$$

Substituting Lemma 2.1, $\sigma_\rho^2(\pi) = \kappa^2(1-\pi)^2 + O(\kappa^3)$, so $\tau^2(\pi) = \log(1+\sigma_\rho^2(\pi)) = \sigma_\rho^2(\pi) + O\big(\sigma_\rho^4(\pi)\big)$. This gives the near-on-policy asymptotic

$$\mathbb{E}[G \mid \pi] \;=\; \mu(\pi)(1-\pi)\,\sigma_\rho^2(\pi) \;+\; O\big(\sigma_\rho^4(\pi)\big) \;=\; \mu(\pi)(1-\pi)^3\,\kappa^2 \;+\; O(\kappa^3), \tag{42}$$

where the second equality substitutes Lemma 2.1 for explicit π-scaling. This recovers a clean "$(1-\pi)^3$ scaling" regime used in §C.3 to characterize the small-$\bar\sigma_\rho^2$ behavior of D.

C.2.6 Large-τ behavior

For large τ, the endpoints satisfy

$$U(\pi) - \tau \;=\; \frac{\log(1+\epsilon_h)}{\tau} - \frac{\tau}{2} \;\to\; -\infty, \qquad L'(\pi) - \tau \;\le\; -\tau \;\to\; -\infty, \tag{43}$$

so $\Phi(U-\tau) - \Phi(L'-\tau) \to 0$ and $H(\tau) \to 0$. Combined with the small-τ growth $H(\tau) \approx \tau^2$, H is unimodal in τ for typical ϵ_h, ϵ_l, peaks at an intermediate τ, and eventually decreases as $\tau \to \infty$. Translating to the local off-policy axis, $\tau^2(\pi) = \log(1+\sigma_\rho^2(\pi))$ grows only logarithmically in $\sigma_\rho^2(\pi)$, so the suppression of H as a function of $\sigma_\rho^2(\pi)$ (and hence of $\bar\sigma_\rho^2$) is polynomial, not exponential.

C.2.7 Numerical illustration setup

The closed form Eq. (39) is fully specified once $(\mu(\pi), p(\pi), \epsilon_h, \epsilon_l)$ and the off-policy degree (parametrized by κ via Lemma 2.1, or equivalently by $\bar\sigma_\rho^2 = \kappa^2\,\mathbb{E}_{p(\pi)}[(1-\pi)^2] + O(\kappa^3)$) are chosen. For the numerical illustrations referenced in this paper, we use clip parameters $(\epsilon_h, \epsilon_l) = (0.2, 0.2)$; representative choices for $p(\pi)$ are taken from the empirical token-probability histogram of a Qwen2.5-7B rollout on math-reasoning prompts; and $\mu(\pi)$ is treated as a slowly-varying non-negative function (we use $\mu(\pi) \equiv \mu_0$ as the simplest representative, which already exhibits the qualitative behavior described in §C.3). These choices are intended for transparency, not parameter fitting; the qualitative phenomena ($H(\tau)$ unimodality, polynomial suppression in $\bar\sigma_\rho^2$, and the dominance reversal in §C.3) are not sensitive to them.

C.2.8 Finite-η caveat

The closed form Eq. (39) is derived under (M3) ($\eta \equiv 0$), which makes $\mathrm{sign}(A) = \mathrm{sign}(u)$ and collapses the trust-region indicator to a region in u alone. With finite η, $\mathrm{sign}(A)$ and $\mathrm{sign}(u)$ can disagree, and Eq. (29) is replaced by truncated-normal moments of A given ρ. We do not develop the finite-η extension here and use (M3) only as a tractable analytical guide for the qualitative phenomena above.

C.2.9 Robustness to the IS-ratio distributional choice

The log-normal model in (M1) is a closed-form-friendly choice: it enforces $\rho > 0$ and $\mathbb{E}[\rho \mid \pi] = 1$ while reducing the trust-region integral to a tilted Gaussian. The qualitative conclusions used in this paper, however, do not rely on the specific distributional choice. To make this concrete, we re-derive the relevant quantities under a Gaussian IS-ratio model and verify that the same $\bar{\sigma}_\rho^2$-axis behavior emerges.

Gaussian IS-ratio model. Replace (M1) with

$$\rho \mid \pi \;\sim\; \mathcal{N}\big(1,\, \sigma_\rho^2(\pi)\big), \tag{M1$'$}$$

keeping (M2) (linear coupling) in the form $A = \mu(\pi)(\rho - 1) + \eta$ and (M3) ($\eta \equiv 0$). Under (M3), $\mathrm{sign}(A) = \mathrm{sign}(\rho - 1)$, so the trust-region indicator collapses to $\mathbf{1}\{\rho \in [1 - \epsilon_l,\, 1 + \epsilon_h]\}$.

Standardize $z := (\rho - 1)/\sigma_\rho(\pi) \sim \mathcal{N}(0, 1)$, with endpoints $a := -\epsilon_l/\sigma_\rho(\pi)$, $b := \epsilon_h/\sigma_\rho(\pi)$. Substituting and using the identities $\int_a^b z\phi(z)\,dz = \phi(a) - \phi(b)$ and $\int_a^b z^2\phi(z)\,dz = a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)$,

$$\mathbb{E}[G \mid \pi] \;=\; (1 - \pi)\,\mu(\pi)\,\sigma_\rho^2(\pi)\,\widetilde{H}\big(\sigma_\rho(\pi)\big), \tag{44}$$

where

$$\widetilde{H}(\sigma_\rho) \;:=\; \big[\Phi(b) - \Phi(a)\big] \;+\; \big[a\phi(a) - b\phi(b)\big] \;+\; \frac{1}{\sigma_\rho}\big[\phi(a) - \phi(b)\big]. \tag{45}$$

Equation (44) plays the role of Eq. (39) for the Gaussian model.

Small-$\bar{\sigma}_\rho^2$ scaling. For small $\sigma_\rho(\pi)$, $a \to -\infty$ and $b \to +\infty$, so $\phi(a), \phi(b), a\phi(a), b\phi(b) \to 0$ and $\Phi(b) - \Phi(a) \to 1$. Hence $\widetilde{H}\big(\sigma_\rho(\pi)\big) \to 1$, and

$$\mathbb{E}[G \mid \pi] \;=\; (1 - \pi)\,\mu(\pi)\,\sigma_\rho^2(\pi) + O\big(e^{-c/\sigma_\rho^2}\big) \;=\; \mu(\pi)(1 - \pi)^3\,\kappa^2 + O(\kappa^3), \tag{46}$$

where the second equality uses Lemma 2.1. This matches the leading-order behavior under (M1) (Eq. (42)) exactly, so the (C1) limit in §C.3.2 is identical: $\widetilde{C}_0$ and C_0 are the same structural constants under either distributional choice.

Large-σ_ρ behavior. For large $\sigma_\rho(\pi)$, $a, b \to 0$, so $\Phi(b) - \Phi(a) \to 0$, $a\phi(a) - b\phi(b) \to 0$, and $\phi(a) - \phi(b) = O(\epsilon_h + \epsilon_l)/\sigma_\rho$. Hence $\widetilde{H} \to 0$ and $\mathbb{E}[G \mid \pi] \to 0$, giving the same trust-region suppression mechanism that drives (C2). The Gaussian model has the well-known artifact that ρ can take negative values when $\sigma_\rho^2(\pi)$ is comparable to 1, which is the principal practical reason we adopt log-normal; the IVT argument and the dominance reversal, however, do not rely on positivity of ρ.

Takeaway. The same closed-form factorization $\mathbb{E}[G \mid \pi] = (1 - \pi)\mu(\pi)f\big(\sigma_\rho(\pi)\big)$, the same near-on-policy $(1 - \pi)^3\,\kappa^2$ scaling, the same C_0, and the same large-σ_ρ suppression hold under either (M1) or (M1$'$). The dominance reversal in §C.3 therefore does not rely on the log-normal choice; (M1) is selected to keep the closed form clean (especially for $\rho > 0$), not as a load-bearing assumption.

C.3 Derivation of Corollary 2.3

This appendix derives the sufficient condition for the gradient dominance reversal stated in Corollary 2.3, using the closed form from Appendix C.2.

C.3.1 Setup

By Proposition 2.2,

$$\mathbb{E}[G \mid \pi] \;=\; (1-\pi)\,\mu(\pi)\,H\big(\tau(\pi)\big), \qquad \tau^2(\pi) \;=\; \log\big(1+\sigma_\rho^2(\pi)\big), \tag{47}$$

with $H(\tau)$ defined in Eq. (39). Fix $0 < p_L < p_H < 1$ and let $I_L = [0, p_L]$, $I_H = [p_H, 1)$. Following the main text, we define the group-level conditional averages with respect to the rollout token-probability density $p(\pi)$:

$$\bar{G}_L \;=\; \mathbb{E}\big[\,\mathbb{E}[G \mid \pi] \,\big|\, \pi \in I_L\big] \;=\; \frac{1}{\mathbb{P}(\pi \in I_L)} \int_0^{p_L} \mathbb{E}[G \mid \pi]\, p(\pi)\, d\pi, \tag{48}$$

$$\bar{G}_H \;=\; \mathbb{E}\big[\,\mathbb{E}[G \mid \pi] \,\big|\, \pi \in I_H\big] \;=\; \frac{1}{\mathbb{P}(\pi \in I_H)} \int_{p_H}^{1} \mathbb{E}[G \mid \pi]\, p(\pi)\, d\pi. \tag{49}$$

We assume $\mathbb{P}(\pi \in I_L), \mathbb{P}(\pi \in I_H) > 0$ throughout.

Off-policy axis. The main-text statement of Corollary 2.3 parametrizes D by the global off-policy degree $\bar{\sigma}_\rho^2 := \mathbb{E}_{p(\pi)}[\sigma_\rho^2(\pi)]$. To carry out the analysis, we use Lemma 2.1, $\sigma_\rho^2(\pi) = \kappa^2(1-\pi)^2 + O(\kappa^3)$, which gives

$$\bar{\sigma}_\rho^2 \;=\; \kappa^2\, c_p \;+\; O(\kappa^3), \qquad c_p := \mathbb{E}_{p(\pi)}\big[(1-\pi)^2\big] \;>\; 0. \tag{50}$$

Hence $\bar{\sigma}_\rho^2$ is a strictly increasing function of κ^2 for small κ, and the IVT statement on $\bar{\sigma}_\rho^2$ is equivalent to the corresponding IVT statement on κ. In what follows we carry out the proof in the natural parametrization κ and translate the conclusions back to $\bar{\sigma}_\rho^2$. Define the difference

$$D(\kappa) \;:=\; \bar{G}_L\big|_\kappa - \bar{G}_H\big|_\kappa, \qquad D(\bar{\sigma}_\rho^2) \;=\; D(\kappa(\bar{\sigma}_\rho^2)), \tag{51}$$

where $\kappa(\bar{\sigma}_\rho^2) = \sqrt{\bar{\sigma}_\rho^2/c_p} + O(\bar{\sigma}_\rho^2)$ inverts Eq. (50).

C.3.2 (C1) Small-$\bar{\sigma}_\rho^2$ behavior of D

By Eq. (42), for each fixed $\pi \in (0,1)$,

$$\mathbb{E}[G \mid \pi] \;=\; \mu(\pi)(1-\pi)^3\,\kappa^2 \;+\; r(\pi,\kappa), \qquad r(\pi,\kappa) = O(\kappa^3), \tag{52}$$

uniformly on compact subsets of $(0,1)$. Integrating against $p(\pi)$ over I_L and I_H,

$$\bar{G}_L\big|_\kappa \;=\; \kappa^2\,\mathbb{E}\big[\mu(\pi)(1-\pi)^3 \,\big|\, \pi \in I_L\big] \;+\; O(\kappa^3), \tag{53}$$

and similarly for $\bar{G}_H\big|_\kappa$. Subtracting,

$$\frac{D(\kappa)}{\kappa^2} \;\xrightarrow[\kappa\to 0^+]{}\; \widetilde{C}_0 \;:=\; \mathbb{E}\big[\mu(\pi)(1-\pi)^3 \,\big|\, \pi \in I_L\big] \;-\; \mathbb{E}\big[\mu(\pi)(1-\pi)^3 \,\big|\, \pi \in I_H\big]. \tag{54}$$

By Eq. (50), $\kappa^2 = \bar{\sigma}_\rho^2/c_p + O\big((\bar{\sigma}_\rho^2)^{3/2}\big)$, so dividing by $\bar{\sigma}_\rho^2$ instead gives

$$\frac{D(\bar{\sigma}_\rho^2)}{\bar{\sigma}_\rho^2} \;\xrightarrow[\bar{\sigma}_\rho^2\to 0^+]{}\; C_0 \;:=\; \frac{\widetilde{C}_0}{c_p}. \tag{55}$$

(C1) is the assumption $C_0 > 0$, equivalently $\widetilde{C}_0 > 0$. Since C_0 is proportional to $\mathbb{E}[\mu(\pi)(1-\pi)^3 \mid I_L] - \mathbb{E}[\mu(\pi)(1-\pi)^3 \mid I_H]$, (C1) holds if and only if the first conditional expectation strictly exceeds the second. Since $(1-\pi)^3$ is strictly larger on I_L than on I_H (as $p_L < p_H$ implies $\inf_{I_L}(1-\pi) > \sup_{I_H}(1-\pi)$), this holds in particular whenever $\mu(\pi) \equiv \mu_0 > 0$ is constant, or more generally whenever $\mu(\pi)$ does not vary so strongly across I_L and I_H as to overcome the $(1-\pi)^3$ advantage. In our numerical illustrations (with $\mu \equiv \mu_0 > 0$ constant), (C1) holds trivially.

Consequence of (C1). By Eq. (55), there exists $s_0 > 0$ such that for all $\bar{\sigma}_\rho^2 \in (0, s_0)$, $D(\bar{\sigma}_\rho^2) > \frac{C_0}{2}\bar{\sigma}_\rho^2 > 0$. In particular, picking any $s_a \in (0, s_0)$ gives $D(s_a) > 0$.

C.3.3 (C2) Existence of an off-policy regime where high-probability tokens dominate

Condition (C2) of Corollary 2.3 asserts that there exists $s_1 > 0$ with $D(s_1) < 0$. (C2) is not a consequence of the modeling choices alone — one can construct degenerate (μ, p) pairs (e.g., $\mu(\pi) = 0$ on I_H, so that I_H never contributes) for which $D(\bar{\sigma}_\rho^2) > 0$ at every $\bar{\sigma}_\rho^2$. We discuss the mechanism that produces $D(s_1) < 0$ for non-degenerate (μ, p), and how (C2) is verified from the closed form in our representative setup.

Mechanism. Eq. (39) shows that the π-dependence of $\mathbb{E}[G \mid \pi]$ enters through $\tau^2(\pi) = \log(1 + \sigma_\rho^2(\pi))$ and through the standardized window $[L'(\pi), U(\pi)]$. Recall from §C.2 that $H(\tau)$ peaks at an intermediate τ and decays for large τ (the trust-region indicator pushes mass into the tail of the tilted log-normal). Since $\sigma_\rho^2(\pi)$ is monotonically decreasing in π, $H(\tau(\pi))$ is maximized at small π for small $\bar{\sigma}_\rho^2$ but for sufficiently large $\bar{\sigma}_\rho^2$ shifts past the peak on I_L first, while $H(\tau(\pi))$ on I_H stays closer to the small-τ regime $H(\tau) \approx \tau^2$. As $\bar{\sigma}_\rho^2$ grows, this mismatch reduces $\bar{G}_L$ relative to $\bar{G}_H$.

Numerical verification. Whether the mismatch is large enough to produce $D(s_1) < 0$ depends on $(p, \mu, \epsilon_h, \epsilon_l)$. Given any concrete choice of these, D is a one-dimensional function of $\bar{\sigma}_\rho^2$ (equivalently κ via Eq. (50)) that can be evaluated by direct numerical integration of Eq. (39) against $p(\pi)$. Using the representative setup of Appendix C.2.7 (i.e., $(\epsilon_h, \epsilon_l) = (0.2, 0.2)$, $\mu(\pi) \equiv \mu_0$, and $p(\pi)$ from a Qwen2.5-7B rollout), we verify numerically that D is positive for small $\bar{\sigma}_\rho^2$ and becomes negative at sufficiently large $\bar{\sigma}_\rho^2$, confirming (C2) for this representative case. The qualitative pattern is robust to mild changes in $p(\pi)$ and to slowly-varying $\mu(\pi)$.

Why the reversal is not exponential. Since $\tau^2(\pi) = \log\big(1 + \sigma_\rho^2(\pi)\big)$ grows only logarithmically in $\sigma_\rho^2(\pi)$, the suppression of $H(\tau(\pi))$ on I_L as $\bar{\sigma}_\rho^2 \to \infty$ is polynomial in $\bar{\sigma}_\rho^2$, not exponential. Consequently, the reversal threshold $s^\star$ may sit at moderate values of $\bar{\sigma}_\rho^2$ rather than in an extreme tail; this is consistent with the empirical pattern in Figure 2(b), where dominance shifts occur over a moderate range of off-policy strengths.

C.3.4 IVT crossing

D is continuous in $\bar{\sigma}_\rho^2 \in (0, \infty)$: $\tau(\pi)$ is continuous in $\sigma_\rho^2(\pi)$ (and hence in $\bar{\sigma}_\rho^2$ via Eq. (50)) for each π, $H(\tau)$ is continuous in τ, and the integrals defining $\bar{G}_L, \bar{G}_H$ are bounded (since H is bounded on bounded τ-ranges and $p(\pi), \mu(\pi), (1 - \pi)$ are bounded on $[0, 1]$), so dominated convergence ensures continuity.

By (C1), there exists $s_a \in (0, s_1)$ with $D(s_a) > 0$.[5] By (C2), $D(s_1) < 0$. By the intermediate value theorem applied to D on $[s_a, s_1]$, there exists at least one $s^\star \in (s_a, s_1)$ with

$$D(s^\star) \;=\; 0. \tag{56}$$

We call any such $s^\star$ a *gradient dominance reversal threshold*. When D changes sign at $s^\star$ (the generic case), low-probability tokens dominate the conditional gradient signal in some left neighborhood of $s^\star$ and high-probability tokens dominate in some right neighborhood. Without further regularity, the IVT alone gives existence of a crossing but not uniqueness or monotonicity — multiple crossings are not excluded by (C1)–(C2). We do not need uniqueness for the qualitative reversal claim made in the main text.

C.3.5 Connection to the empirical observation

The empirical pattern in Figure 2(b) is consistent with this model-level prediction. Two caveats: (i) the figure measures parameter-gradient norms aggregated over tokens, not the closed-form $D(\bar{\sigma}_\rho^2)$ directly; (ii) the model's $\bar{\sigma}_\rho^2$ summarizes off-policy strength, while the experimental degree of off-policy is controlled indirectly via the number of updates per rollout. The match should therefore be read as evidence that the qualitative dominance reversal predicted by the closed form is observable in practice, rather than as a quantitative validation of the closed form itself.

[5] Take any $s_a \in (0, \min(s_0, s_1))$ where s_0 is from §C.3.2.

D Geometric Direction Analysis with Principal Angles

This appendix details the geometric tool used in Section 3.1 to quantify the *directional* similarity between high-dimensional gradients. Our goal is to compare gradients in a way that is (i) robust to global rescaling, (ii) stable in very high dimensions, and (iii) compatible with the matrix-shaped structure of transformer layer gradients.

D.1 Why subspace alignment instead of vector angles

A naive approach is to flatten all parameters and compute a cosine similarity between two gradient vectors. However, for large models this is often dominated by noise and by the many low-energy directions. In addition, gradients in neural networks are typically *low-rank structured*: most update energy concentrates in a relatively small set of directions (e.g., due to correlated features and shared activations). For these reasons, we compare the *dominant gradient subspaces* rather than individual flattened vectors.

Concretely, given a layer parameter matrix (e.g., a linear projection weight) with gradient

$$W \in \mathbb{R}^{m\times n},$$

we treat W as a linear operator and extract its top-k right singular vectors as a compact representation of its dominant update directions.

D.2 Subspace representation via SVD

For each gradient matrix W, we compute the singular value decomposition (SVD):

$$W = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthonormal matrices, and Σ is diagonal with singular values in non-increasing order.

We define the rank-k dominant *right* singular subspace

$$\mathcal{S}(W;k) \;=\; \mathrm{span}(V_k),$$

where $V_k \in \mathbb{R}^{n\times k}$ contains the top-k columns of V. Intuitively, V_k captures the principal directions in the input-feature space along which the gradient operator varies most strongly.

Choice of right singular vectors. We use right singular vectors because for weight matrices mapping $\mathbb{R}^n \to \mathbb{R}^m$, V_k corresponds to directions in the input dimension. Empirically we found V_k yields stable comparisons across layers of similar shapes. (Using left singular vectors U_k gives qualitatively similar conclusions for our use case.)

Choosing k. We set $k = 128$ in all experiments unless otherwise noted. This value is large enough to capture most gradient energy for typical transformer projection matrices, while keeping computations tractable. In preliminary sweeps, we found the mean principal angle is qualitatively stable for k in a broad range (e.g., 64–256), provided k does not exceed the smaller matrix dimension.

D.3 Principal angles between gradient subspaces

Given two gradients W^A and W^B for the same layer (e.g., computed under two different update constructions), we form orthonormal bases for their dominant subspaces:

$$Q^A \in \mathbb{R}^{n\times k}, \quad Q^B \in \mathbb{R}^{n\times k},$$

where $Q^A = V_k(W^A)$ and $Q^B = V_k(W^B)$.

The k principal angles $\{\theta_i\}_{i=1}^k$ between the two subspaces are defined by:

$$\cos(\theta_i) \;=\; s_i,$$

where $\{s_i\}_{i=1}^k$ are the singular values of the cross-Gram matrix

$$M \;=\; (Q^A)^\top Q^B \in \mathbb{R}^{k\times k}.$$

Equivalently, if $M = \tilde{U}\,\mathrm{diag}(s_1,\ldots,s_k)\,\tilde{V}^\top$, then

$$\theta_i = \arccos(s_i), \quad i = 1,\ldots,k.$$

Interpretation. Each $\theta_i \in [0°, 90°]$ measures how well the i-th most-alignable direction in one subspace matches the other. In particular:

- $\theta_i = 0°$ indicates a perfectly shared direction.
- $\theta_i = 90°$ indicates orthogonality for that principal direction.

Small angles across many i imply the two gradients share a similar dominant update subspace; large angles indicate a substantial rotation of dominant directions.

D.4 Summary statistics reported

The set of principal angles yields a spectrum of directional similarity. For compact reporting, we aggregate angles into summary statistics.

Mean principal angle. We report the mean principal angle:

$$\bar{\theta}(W^A, W^B) \;=\; \frac{1}{k}\sum_{i=1}^{k}\theta_i.$$

This provides a single scalar notion of subspace alignment. In our plots/tables, smaller $\bar{\theta}$ indicates stronger alignment.

Alternative aggregations (not required). Other useful summaries include the median, the maximum angle, or a weighted mean using singular values of W as weights. We found the unweighted mean sufficient and stable for the comparisons in this paper.

D.5 Numerical stability

In exact arithmetic, all singular values of $M = (Q^A)^\top Q^B$ lie in $[0, 1]$. In finite precision, slight violations (e.g., $1 + 10^{-7}$) can occur, which would make $\arccos(\cdot)$ undefined. We therefore clamp:

$$s_i \leftarrow \min(1, \max(0, s_i)),$$

before applying $\arccos$.

D.6 Layer-wise computation and aggregation across layers

We compute principal angles *per layer* and (when presenting model-level summaries) aggregate across layers.

Per-layer procedure. For each layer weight matrix:

1. Compute gradients W^A and W^B under two update constructions.
2. Compute truncated SVDs to obtain Q^A and Q^B.
3. Form $M = (Q^A)^\top Q^B$, compute its singular values $\{s_i\}$, then angles $\{\theta_i\}$.
4. Summarize with $\bar{\theta}$ (and optionally the full spectrum).

Across-layer aggregation. To summarize directional similarity for a whole model, we average $\bar{\theta}$ over a set of layers. In the main experiments we include the main projection matrices (e.g., attention and MLP projections) and exclude bias vectors and layernorm parameters, for which matrix-shaped SVD-based comparisons are not meaningful.

D.7 Practical notes

Truncated SVD implementation. We use a truncated SVD to compute the top-k singular vectors efficiently. For large matrices, randomized SVD yields significant speedups while maintaining sufficient accuracy for principal-angle estimates.

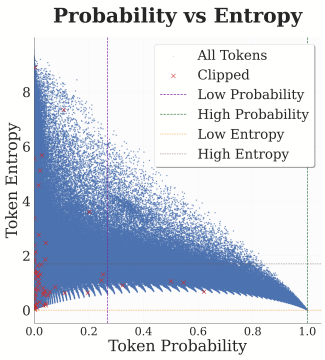

Figure 5: Correlation between token probability and entropy, with clipped tokens highlighted (red crosses).

Why compare subspaces, not singular vectors directly. Singular vectors are defined up to sign flips and may swap under small perturbations when singular values are close. Subspace-based measures are invariant to such instabilities, making principal angles more reliable for comparing gradients from stochastic minibatches.

D.8 Limitations

Principal angles measure similarity of dominant *linear* subspaces per layer. This is not a full characterization of optimization trajectories, and it does not capture higher-order geometry or parameter coupling across layers. Nevertheless, it is a simple, robust diagnostic for whether two update constructions point in broadly similar directions in parameter space.

D.9 Reproducibility checklist

For reproducibility, we specify:

- Subspace dimension: $k = 128$.
- Angle computation: SVD of $(Q^A)^\top Q^B$ with clamping to $[0, 1]$.
- Layer selection: matrix-shaped parameters (projection weights); exclude biases and layer-norm scalars.
- Reported statistic: mean principal angle $\bar{\theta}$ (and spectra in some ablations).

E Reconciling Heuristics: Probability, Entropy, and Clipping Dynamics

The empirical contradiction is resolved by examining the interplay between probability, entropy, and the clipping mechanism. While a token's probability π and the vocabulary's entropy H are correlated, their formal relationship is bounded:

$$\underbrace{-\pi \log \pi - (1-\pi)\log(1-\pi)}_{H_{\min}(\pi)} \leq H \leq \underbrace{-\pi \log \pi - (1-\pi)\log\left(\frac{1-\pi}{V-1}\right)}_{H_{\max}(\pi)}$$

Figure 5 visualizes these bounds and provides the crucial insight: clipped tokens (red crosses) are not randomly distributed but are overwhelmingly concentrated in the low-probability, low-entropy region.

This single observation directly reconciles the conflicting findings:

- Updating **low-probability tokens** [Yang et al., 2025b] is suboptimal in the near on-policy regime because this high-variance population then dominates the update (Section 2). As

established in Lemma 2.1, these tokens also have the widest IS-ratio spread, so in more off-policy regimes they are the most susceptible to clipping, which attenuates their gradient contribution.

- Conversely, updating **high-entropy tokens** [Wang et al., 2025b] succeeds because it acts as an effective filter for stability. This heuristic implicitly avoids the heavily clipped region, selecting for tokens from more uniform distributions that are robust to policy updates.

Therefore, entropy is not merely a proxy for low probability but a more discerning indicator of a token's dynamic stability under the clipping pressures that govern off-policy optimization.

F Detailed Gradient Analysis

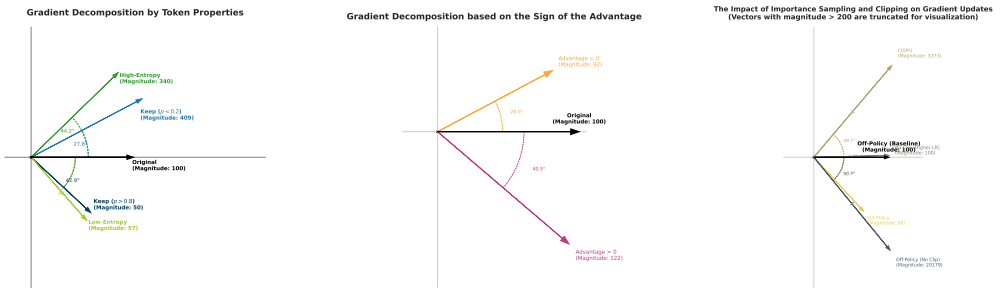

Figure 6: Schematic illustration of the impact of different factors on RL updates. Reported angles are mean principal angles between dominant gradient subspaces (Appendix D), not geometric angles between the schematic arrows.

F.1 Token Property Analysis

To understand the fundamental composition of the policy gradient, we analyze the contributions of different token subsets based on their intrinsic properties. Our analysis, presented in Figure 6 left (measured under near on-policy training), isolates the gradient signals originating from distinct populations of tokens, revealing that the update is overwhelmingly dominated by a small, low-probability subset. Consistent with Yang et al. [2025b], tokens with $\pi_k < 0.2$ produce a gradient with a norm 409% that of the full gradient and are exceptionally well-aligned with its direction (27.8° mean principal angle). In stark contrast, high-probability tokens ($\pi_k > 0.8$) contribute a much smaller, less aligned gradient (50% norm, 42.9° angle). This confirms that the policy gradient is primarily driven by powerful but potentially unstable signals from the tail of the probability distribution.

The strategy of updating high-entropy tokens, proposed by Wang et al. [2025b], can be understood as an effective heuristic for managing this instability. The high-entropy subset produces a gradient with a substantial magnitude (340% of original) but a weaker directional alignment (44.2°) than the pure low-probability set. This suggests the high-entropy criterion functions as a filter: it selects for tokens that are informative (often having low to moderate probability) but originate from flatter, more uncertain distributions. As we will show, these flatter distributions are inherently more stable under policy updates. Therefore, the heuristic succeeds by implicitly balancing gradient magnitude against stability. This success underscores the need for a more principled mechanism that can dynamically manage this trade-off.

F.2 Advantage Sign Analysis

The advantage function, A, guides learning by signaling whether an action should be reinforced ($A > 0$) or suppressed ($A < 0$). To understand their distinct roles, we analyze the gradient contributions from these two subsets separately, as shown in Figure 6 middle. Our analysis reveals a critical asymmetry: while the gradient norms from positive (122% of original) and negative (92%)

advantage samples are comparable, their directional guidance differs markedly. The gradient from negative-advantage samples is substantially better aligned with the total gradient direction (28.0° mean principal angle) than that from positive-advantage samples (40.5°). This indicates that the overall update direction is predominantly dictated by corrective signals from suboptimal actions.

This directional dominance stems from the fundamentally different effects of positive and negative updates. An update with $A > 0$ reinforces a single action, implicitly suppressing all alternatives and potentially narrowing the policy. Conversely, an update with $A < 0$ penalizes a specific action, which effectively redistributes probability mass across the rest of the vocabulary. This latter mechanism, a form of error correction, promotes exploration and enhances policy diversity. The geometric dominance of negative-advantage gradients therefore suggests that learning in complex tasks is driven more by correcting errors than by reinforcing known correct pathways, a process crucial for discovering robust strategies. While Zhu et al. [2025] also noted the importance of negative samples, our work provides a novel perspective by demonstrating their dominant role in shaping the gradient's geometric direction.

F.3 Importance Sampling Analysis

To improve sample efficiency, off-policy reinforcement learning corrects for the policy distribution mismatch using Importance Sampling (IS). However, the high variance of the IS ratio necessitates a stabilization mechanism like clipping, which modulates the raw gradient into a stable update signal. To dissect this process, we compare our baseline off-policy GRPO update against several variants (Figure 6 right). The indispensability of clipping is starkly illustrated by its removal: the gradient norm explodes to **20,179%** of the baseline, and its direction severely deviates (**50.7**° mean principal angle). This confirms clipping is crucial not just for controlling magnitude but also for maintaining a stable optimization path.

Perhaps the most revealing finding is the significant directional divergence (**47.4**°) between the on-policy and standard off-policy (GRPO) gradients, despite their nearly identical magnitudes (98% vs. 100%). This highlights a fundamental trade-off: to maintain magnitude stability, GRPO's fixed clipping mechanism systematically filters out certain updates (predominantly from high-variance, low-probability tokens), creating a gradient direction that is geometrically distinct from the on-policy ideal. This implies that while off-policy learning is sample-efficient, its optimization trajectory can substantially differ from its on-policy counterpart.

The specific design of the clipping strategy further modulates this trade-off. For instance, CISPO, which clamps the IS ratio instead of gating the entire gradient to zero, retains more signal from outlier tokens. This results in a substantially larger gradient norm (**3,273%**) and a similarly large directional shift (**49.7**°), reflecting a more aggressive update policy. These comparisons reveal that the standard clipping in methods like PPO/GRPO is not a neutral stabilizer; it actively shapes the gradient by discarding certain information. This suggests an opportunity for adaptive mechanisms that can better preserve the on-policy direction while retaining off-policy efficiency.

G ACPO Formulation (Details)

ACPO replaces the single global clipping range in Eq. 1 with a bin-specific range determined by the behavior-token probability. Given an update mini-batch (pooling all tokens in the update batch), we assign each token (i, t) to one of B equal-width bins on $[0, 1]$ via

$$b(i,t) = \min\{B, \lfloor B \cdot \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t}) \rfloor + 1\}. \tag{57}$$

For each bin b, we compute the within-bin dispersion of IS ratios on this mini-batch (a stop-gradient statistic, recomputed independently for every optimization mini-batch; see Appendix N.4) and set the clipping range, clamped to $[\epsilon_{\min}, \epsilon_{\max}]$, as

$$\begin{aligned} \sigma_b &= \mathrm{Std}\big(\{\rho_{i,t} : b(i,t) = b\}\big), \\ \epsilon_b &= \mathrm{clip}\big(\epsilon_{\text{base}} + \alpha\, \mathrm{sg}[\sigma_b],\, \epsilon_{\min},\, \epsilon_{\max}\big). \end{aligned} \tag{58}$$

Algorithm 1 ACPO (bin-wise adaptive clipping)

Input: behavior policy $\pi_{\theta_{\text{old}}}$, current policy π_θ, batch of rollouts, B, ϵ_{base}, α.
Compute token-wise IS ratios $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$.
Assign each token (i,t) to a probability bin $b(i,t) = \min\{B, \lfloor B \cdot \pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})\rfloor + 1\}$.
for $b = 1, \ldots, B$ **do**
    $\sigma_b \leftarrow \text{Std}(\{\rho_{i,t} : b(i,t) = b\})$.
    $\epsilon_b \leftarrow \text{clip}(\epsilon_{\text{base}} + \alpha\, \text{sg}[\sigma_b],\, \epsilon_{\min},\, \epsilon_{\max})$.
end for
Optimize Eq. 59 using token-wise clipping bounds $(1 - \epsilon_{b(i,t)},\, 1 + \epsilon_{b(i,t)})$.

We then apply token-wise clipping by using $\epsilon_{b(i,t)}$ in Eq. 1. The resulting objective is

$$J_{\text{ACPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big[\rho_{i,t}A_i,\ \text{clip}\big(\rho_{i,t}, 1-\epsilon_{b(i,t)}, 1+\epsilon_{b(i,t)}\big)A_i\Big]. \tag{59}$$

H More Related Works

Reinforcement Learning for Large Language Models. Reinforcement learning has evolved from a tool for preference alignment to a key technique for enhancing reasoning capabilities in large language models. Initially pioneered through Reinforcement Learning from Human Feedback (RLHF), RL methods like PPO [Schulman et al., 2017] were used to align models with human preferences using human-annotated data [Ouyang et al., 2022]. The landscape shifted dramatically with the emergence of RL with verifiable rewards (RLVR), which leverages objective, automatically verifiable feedback signals instead of subjective human preferences [Lambert et al., 2024]. OpenAI's o1 model [Jaech et al., 2024] first showcased that RLVR can effectively incentivize reasoning at scale, particularly in tasks like mathematics and programming. Building on this foundation, researchers developed improved algorithmic methods such as GRPO [Shao et al., 2024] and its variants (e.g., DAPO [Yu et al., 2025]). Subsequent models trained with these methods, including DeepSeek-R1 [Guo et al., 2025a], QwQ [Team, 2025], and AceReason-Nemotron [Chen et al., 2025c], demonstrate that strong reasoning capabilities can emerge through outcome-based optimization with online RL algorithms, establishing RLVR as a promising paradigm for developing reasoning-capable LLMs.

Token-Level Dynamics and Update Strategies. While RLVR has demonstrated strong potential, its effectiveness is often determined by how token-level updates are applied during training. Recent research has explored multiple related directions: entropy- or probability-based methods that incorporate or reduce uncertainty [Wang et al., 2025b, Yang et al., 2025b, Cheng et al., 2025, Cui et al., 2025, Gao et al., 2025, Chen et al., 2025b], modifications to importance-sampling and clipping that stabilize gradients [Roux et al., 2025, Chen et al., 2025a, Su et al., 2025, Zheng et al., 2025a, Wang et al., 2025a], advantage design strategies ranging from negative reinforcement to minimalist or segment-level credit assignment [Zhu et al., 2025, Xiong et al., 2025], and off-policy approaches that reshape update distributions [Yan et al., 2025, Ma et al., 2025, Fu et al., 2025, Arnal et al., 2025]. However, studies in this area sometimes yield conflicting conclusions: for example, Wang et al. [2025b] emphasize that high-entropy minority tokens drive effective learning, whereas Yang et al. [2025b] show that low-probability (usually high-entropy) tokens can over-dominate gradients and should be suppressed. Such inconsistencies call for a deeper, component-wise analysis of the RL update pipeline. While recent work has started to examine how individual training techniques affect RL dynamics [Liu et al., 2025], we still lack a comprehensive framework that can systematically explain how these components work together to drive effective policy updates in LLM reasoning.

I Training Details

We conduct experiments using the Qwen2.5-7B model on the ORZ dataset. All methods are evaluated under two settings: near on-policy (N-OnP) with a train batch size of 128 and PPO mini batch size of 64, and off-policy (OffP) with a train batch size of 256 and PPO mini batch size of 16. The key

Table 3: Common hyperparameters for Qwen2.5-7B experiments on ORZ-57K (math reasoning).

Hyperparameter	**N-OnP**	**OffP**
Base Model	Qwen2.5-7B	
Max Prompt Length	1024	
Max Response Length	3072	
Learning Rate	1e-6	
Total Epochs	15 (budget; stopped at $\sim$1,000 steps)	
Rollout N	10	
Temperature	1.0	
Validation N	8	
ϵ_{low}	0.2	
ϵ_{high}	0.3	
Entropy Coefficient	0.0	
KL Coefficient	0.0	
Loss Aggregation	seq-mean-token-mean	
Train Batch Size	128	256
PPO Mini Batch Size	64	16

difference between these settings is the number of optimizer updates performed per rollout: 2 in the near on-policy setting and 16 in the off-policy setting. The latter reuses each rollout for more gradient steps before the behavior-policy samples are refreshed, resulting in greater policy staleness.

All methods share the following common hyperparameters: a learning rate of 1e-6, a maximum prompt length of 1024 tokens, a maximum response length of 3072 tokens, and a rollout size $N = 10$. The training budget is set to 15 epochs, but all runs are stopped after about 1,000 steps. We use no KL regularization (KL coefficient = 0) and no entropy bonus (entropy coefficient = 0). Our training code is adapted from the VeRL framework [Sheng et al., 2024]. Settings not listed here (e.g., the optimizer configuration and the Countdown and HiTab runs) follow the configuration files in our released code repository (`https://github.com/Control-derek/ACPO`). The loss aggregation strategy is seq-mean-token-mean, which first averages over tokens within each sequence and then averages across sequences.

DAPO. The baseline method uses asymmetric clipping with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.3$, without any additional modifications.

AR-Lopti. This method combines Advantage Reweighting (AR) and Low-Probability Token Isolation (Lopti). The AR component reweights advantages using $\alpha_{\text{AR}} = 0.3$, $\tau_{\text{AR}} = 0.7$, and negative advantage weight $w_{\text{neg}} = 1.0$. The Lopti component isolates low-probability tokens using a threshold $p_{\text{lopti}} = 0.5$.

High-Entropy. This method selectively updates only the tokens with the highest entropy. We set $k_{\text{ratio}} = 0.2$, meaning only the top 20% highest-entropy tokens are updated during training.

Low-Entropy. Conversely, this method updates only the tokens with the lowest entropy. We set $k_{\text{ratio}} = 0.8$, meaning only the bottom 80% (i.e., lowest-entropy) tokens are updated.

ACPO. Adaptive Clip Policy Optimization dynamically adjusts the clipping range based on the within-bin importance-sampling ratio variance. We use $\alpha = 3.0$ and $\epsilon_{\text{base}} = 0.2$, with the adaptive clip range bounded between $\epsilon_{\text{min}} = 0.0$ and $\epsilon_{\text{max}} = 3.0$.

CISPO. This method applies importance sampling-aware clipping with separate clip ratios for the importance sampling ratio: $\epsilon_{\text{is_high}} = 0.45$ and $\epsilon_{\text{is_low}} = 1.0$.

Table 4: Method-specific hyperparameters.

Method	**Specific Hyperparameters**
DAPO	(Baseline, no additional parameters)
AR-Lopti	$\alpha_{\text{AR}} = 0.3$ $\tau_{\text{AR}} = 0.7$ $w_{\text{neg}} = 1.0$ $p_{\text{lopti}} = 0.5$
High-Entropy	entropy_top_k = True $k_{\text{ratio}} = 0.2$
Low-Entropy	entropy_bottom_k = True $k_{\text{ratio}} = 0.8$
ACPO	variance_adaptive_clip = True $\alpha = 3.0$ $\epsilon_{\text{base}} = 0.2$ $\epsilon_{\text{min}} = 0.0$ $\epsilon_{\text{max}} = 3.0$
CISPO	$\epsilon_{\text{is_high}} = 0.45$ $\epsilon_{\text{is_low}} = 1.0$

Table 5: Sensitivity of ACPO to α with fixed $\epsilon_{\text{base}} = 0.2$ on ORZ-trained models, evaluated on multiple mathematical reasoning benchmarks, avg@8. Within each regime, the best score for each benchmark is **bolded**.

α	ϵ_{base}	Minerva	Math500	AMC	AIME24	AIME25	Olympiad	AVG
Near On-Policy (2 updates/rollout)								
2	0.2	**31.23**	**79.85**	53.92	20.00	12.50	42.66	40.03
3	0.2	31.02	79.78	**54.37**	20.83	**13.33**	42.83	**40.36**
4	0.2	30.64	79.63	53.31	**21.25**	12.92	**43.17**	40.15
Off-Policy (16 updates/rollout)								
2	0.2	31.46	80.35	50.90	15.42	10.00	41.67	38.30
3	0.2	**31.66**	**80.48**	**51.20**	16.25	11.25	42.17	38.84
4	0.2	31.51	80.43	50.66	**16.67**	**12.08**	**42.41**	**38.96**

J Ablation Study

J.1 Parameter Sensitivity

We study the sensitivity of ACPO to its two main hyperparameters: the scaling factor α in the adaptive rule $\epsilon_b = \epsilon_{\text{base}} + \alpha\,\sigma_b$, and the base clipping range ϵ_{base}. Tables 5 and 6 report results on ORZ-trained models, evaluated on in-domain and OOD math benchmarks, under both near on-policy (2 updates/rollout) and off-policy (16 updates/rollout) regimes.

Overall, ACPO is reasonably robust across a broad range of settings. In particular, moderate values (e.g., $\alpha \in \{3, 4\}$ with $\epsilon_{\text{base}} \approx 0.2$) achieve consistently strong average performance, while the per-benchmark optima can shift slightly between the on-policy and off-policy regimes, reflecting different clipping dynamics.

J.2 Effects of Clipping Thresholds

A natural question is whether ACPO's improvements stem from the adaptive mechanism itself, or simply from using different clipping threshold values. To disentangle these factors, we conduct ablation experiments on the Countdown task with Qwen2.5-3B-Instruct (Table 7), where we match fixed thresholds to ACPO's observed statistics: (1) **Uniform (Clip=Avg)**: both upper and lower

Table 6: Sensitivity of ACPO to ϵ_{base} with fixed $\alpha = 3$ on ORZ-trained models, evaluated on multiple mathematical reasoning benchmarks, avg@8. Within each regime, the best score for each benchmark is **bolded**.

α	ϵ_{base}	Minerva	Math500	AMC	AIME24	AIME25	Olympiad	AVG
Near On-Policy (2 updates/rollout)								
3	0.0	29.87	77.95	48.04	19.58	**13.75**	41.20	38.40
3	0.1	30.65	78.33	50.90	20.42	13.33	42.10	39.29
3	0.2	**31.02**	**79.78**	**54.37**	20.83	13.33	42.83	**40.36**
3	0.3	30.73	79.23	52.71	**21.25**	12.92	**42.99**	39.97
Off-Policy (16 updates/rollout)								
3	0.0	29.50	75.10	45.78	15.83	10.00	40.30	36.09
3	0.1	30.71	77.73	48.49	16.25	10.83	41.33	37.56
3	0.2	**31.66**	**80.48**	**51.20**	16.25	11.25	42.17	**38.84**
3	0.3	31.12	78.88	50.15	**16.67**	**11.67**	**42.36**	38.48

thresholds set to ACPO's weighted average clip range; (2) **DAPO (Clip-High=Max)**: upper threshold set to ACPO's maximum observed clip range.

Under the *avg* setting, a single threshold equal to ACPO's weighted average clip range does not improve over the DAPO baseline (72.55% vs. 73.27%), since it averages over groups with very different IS-ratio dispersion. Under the *max* setting, matching DAPO's upper bound to ACPO's maximum admits more signals and yields modest gains (74.65% vs. 73.27%). However, this uniform expansion ignores token-level variance: some tokens remain over-clipped while others become overly permissive, leaving performance below ACPO (75.74%).

These results confirm that ACPO's benefit stems from adaptive calibration, rather than simply using larger or smaller threshold values.

J.3 Sensitivity to Number of Bins

We study how ACPO depends on the number of probability bins B used for adaptive thresholding. Table 8 reports results on ORZ-trained Qwen2.5-7B over six math benchmarks under both near on-policy and off-policy regimes, and Table 9 reports Countdown results on Qwen2.5-3B-Instruct.

The default $B = 5$ achieves the best average performance, while $B = 4$ and $B = 6$ remain very close in all settings, indicating that ACPO is robust to moderate changes in bin count.

K Clipping and Gradient Dynamics of ACPO

To understand the mechanisms underlying ACPO's stability, we analyze token-level clipping behavior and gradient dynamics during off-policy training on ORZ. Figure 7 compares DAPO and ACPO across five probability bins.

Clip fraction. DAPO exhibits heterogeneous clipping rates across bins: low-probability tokens (0.0–0.4) are clipped at roughly 20–45%, while high-probability tokens (0.8–1.0) at only about 5–15%. This disparity arises because fixed thresholds cannot accommodate the varying IS-ratio variance—tokens with inherently higher variance exceed the threshold more frequently, leading to over-clipping that discards potentially informative gradient signals. ACPO keeps clip rates lower and more uniform across bins (mostly below about 15%, with one transient spike to about 25% in the

Table 7: Ablation study on Countdown with Qwen2.5-3B-Instruct.

Method / Setting	**N-OnP.**	**OffP.**
DAPO (Baseline)	73.27	74.38
DAPO (Clip-High=Max)	74.65	75.86
Uniform (Clip=Avg)	72.55	74.16
ACPO	**75.74**	**76.27**

Table 8: Bin-number sensitivity of ACPO on ORZ-trained Qwen2.5-7B across math benchmarks. All entries are avg@8.

Regime	Bins	Minerva	Math500	AMC	AIME24	AIME25	Olympiad	Avg.
N-OnP.	4	30.93	78.60	52.56	19.58	12.08	41.86	39.27
	5	31.02	79.78	54.37	20.83	13.33	42.83	40.36
	6	30.58	78.88	53.92	17.92	14.17	41.33	39.47
OffP.	4	30.65	80.28	50.75	17.08	10.42	40.64	38.30
	5	31.66	80.48	51.20	16.25	11.25	42.17	38.84
	6	30.10	79.48	51.81	16.67	9.17	42.43	38.28

Table 9: Bin-number sensitivity on Countdown (Qwen2.5-3B-Instruct).

Bins	Countdown (N-OnP.)	Countdown (OffP.)
4	75.20	75.90
5	75.74	76.27
6	75.43	75.65

0.6–0.8 bin around step 800) by setting per-bin thresholds as $\epsilon_b = \epsilon_{\text{base}} + \alpha \cdot \sigma_b$, which naturally accommodates the heterogeneous variance structure.

Mean gradient. DAPO shows highly unstable gradients, with the mean gradient spiking to about 0.0045 for the 0.2–0.4 bin around step 700. ACPO keeps the mean gradient below about 0.001 throughout training for all bins, confirming that adaptive clipping provides consistent regularization across different token groups and training stages.

L Computational Overhead Analysis

Hardware and wall-clock cost. All training runs were carried out on a single node with 4×NVIDIA A100 80GB GPUs. Under this configuration, a full RL training run takes approximately 14 days for the math reasoning task, 7 days for HiTab, and 2 days for Countdown. Reported numbers correspond to the experiments shown in the paper; preliminary and ablation runs consumed additional compute on the same hardware.

We analyze the computational overhead of different methods by measuring the per-token update time during training. Table 10 summarizes the results across both Off-Policy and Near On-Policy settings.

Why ACPO has negligible overhead. ACPO (Adaptive Clip Policy Optimization) adapts the clipping threshold based on the variance of importance-sampling ratios within each probability bin. The additional computation involves: (1) computing the within-bin standard deviation of IS ratios (a single reduction operation per bin), and (2) adjusting the clip ratio accordingly. Both operations are simple scalar computations performed once per batch, adding negligible overhead to the per-token update time.

Why AR-Lopti has significant overhead. AR-Lopti combines two techniques that both require additional per-token computations:

- **Lopti (Low-Probability Token Isolation)**: Requires computing the probability of each token, comparing against a threshold ($p_{\text{lopti}} = 0.5$), and creating a binary mask to isolate low-probability tokens. This involves additional tensor operations and conditional logic.
- **AR (Advantage Reweighting)**: Applies a non-linear reweighting function to advantages based on their signs and magnitudes, which requires additional computation beyond simple clipping.

Together, these operations result in over 2× computational overhead compared to DAPO.

Why High-Entropy has moderate overhead. The High-Entropy method requires computing the entropy of each token's output distribution, sorting tokens by entropy, and selectively updating only

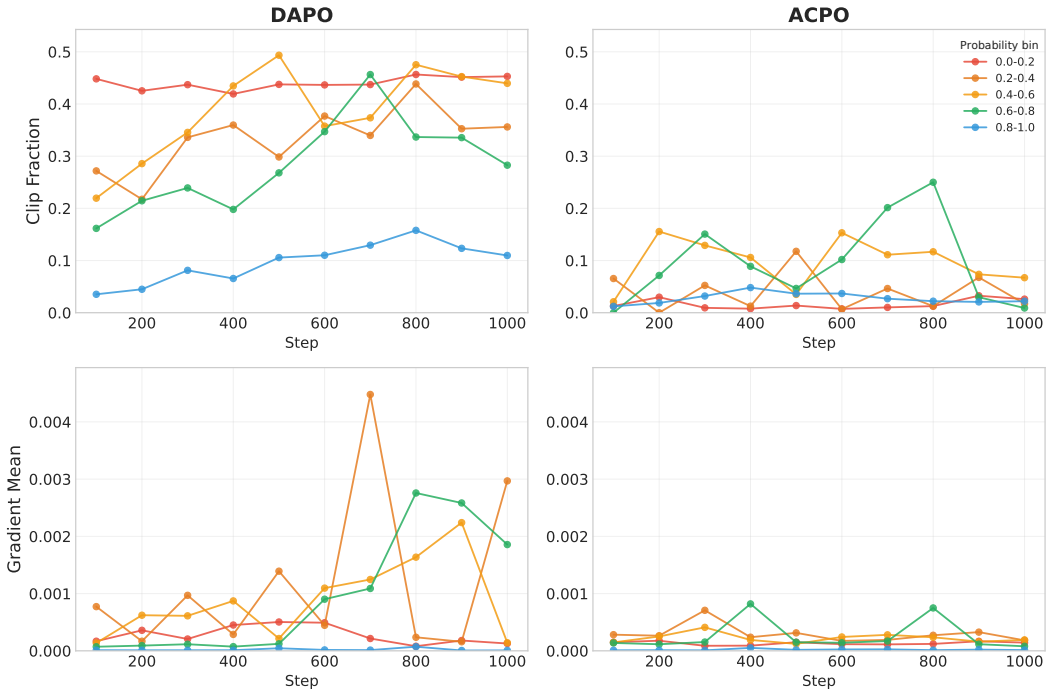

Figure 7: Training dynamics under off-policy training across five probability bins. **Top:** Clip fraction over training steps. DAPO's fixed threshold yields heterogeneous rates; ACPO achieves lower and more uniform rates. **Bottom:** Mean gradient over training steps (per probability bin). DAPO exhibits unstable spikes; ACPO remains stable throughout.

Table 10: Update time comparison. We report both per-step update time (seconds) and per-token update time (milliseconds). ACPO introduces negligible computational overhead compared to DAPO, while AR-Lopti incurs significant overhead due to additional token-level operations.

Method	**N-OnP**				**OffP**			
	Update (s)	Rel.	Per-token (ms)	Rel.	Update (s)	Rel.	Per-token (ms)	Rel.
DAPO	231.0 ± 31.7	–	0.110 ± 0.005	–	440.1 ± 48.9	–	0.118 ± 0.003	–
CISPO	239.6 ± 24.6	+3.8%	0.128 ± 0.006	+15.9%	485.1 ± 49.1	+10.2%	0.136 ± 0.004	+14.7%
AR-Lopti	359.4 ± 32.8	+55.6%	0.234 ± 0.006	+112.3%	659.9 ± 72.5	+49.9%	0.258 ± 0.017	+117.7%
High-Entropy	181.9 ± 22.9	−21.2%	0.114 ± 0.005	+3.3%	556.1 ± 107.6	+26.4%	0.167 ± 0.028	+40.6%
ACPO (Ours)	219.1 ± 45.5	−5.2%	0.110 ± 0.005	+0.0%	458.4 ± 84.6	+4.2%	0.109 ± 0.005	−8.0%

the top-k highest-entropy tokens. The entropy computation and sorting operations add approximately 40% per-token overhead in the off-policy setting (Table 10).

In summary, ACPO achieves its performance improvements through computationally efficient per-bin adaptive clipping, making it practical for large-scale training without additional infrastructure requirements.

M Statistical Significance Tests

We conduct statistical significance tests to compare ACPO against baseline methods on the OOD math benchmarks. For each benchmark, we perform independent two-sample Welch's t-tests (unequal variances) using a one-tailed alternative hypothesis that ACPO outperforms the baseline: H_1 : $\mu_{\text{ACPO}} > \mu_{\text{baseline}}$.

The test statistic is computed as:

$$t = \frac{\bar{X}_{\text{ACPO}} - \bar{X}_{\text{baseline}}}{\sqrt{\frac{s^2_{\text{ACPO}}}{n_{\text{ACPO}}} + \frac{s^2_{\text{baseline}}}{n_{\text{baseline}}}}}, \tag{60}$$

Table 11: One-tailed Welch's t-test p-values comparing ACPO against each baseline on OOD math benchmarks, under near on-policy (2 updates/rollout) and off-policy (16 updates/rollout) settings. Bold indicates $p < 0.05$.

Near On-Policy (2 updates/rollout)						
ACPO vs.	Minerva	Math500	AMC	AIME24	AIME25	Olympiad
Base	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
AR-Lopti	**0.017**	< 0.001	< 0.001	0.052	< 0.001	< 0.001
High-Entropy	**0.018**	< 0.001	**0.029**	**0.002**	< 0.001	< 0.001
Low-Entropy	**0.007**	< 0.001	< 0.001	0.209	< 0.001	< 0.001
DAPO	0.143	**0.002**	**0.007**	–	**0.005**	**0.007**
CISPO	0.285	**0.008**	< 0.001	–	< 0.001	< 0.001
Off-Policy (16 updates/rollout)						
ACPO vs.	Minerva	Math500	AMC	AIME24	AIME25	Olympiad
Base	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
AR-Lopti	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
High-Entropy	**0.005**	< 0.001	< 0.001	**0.013**	–	< 0.001
Low-Entropy	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
DAPO	**0.027**	< 0.001	**0.001**	–	< 0.001	< 0.001
CISPO	**0.007**	< 0.001	**0.003**	**0.002**	**0.02**	< 0.001

where $\bar{X}$ denotes the sample mean, s the sample standard deviation, and n the sample size. The degrees of freedom are approximated using the Welch–Satterthwaite equation. We use $n = 8$ for Minerva, Math500, AMC, and Olympiad benchmarks, and $n = 32$ for AIME24 and AIME25. We consider results statistically significant when $p < 0.05$.

N Additional Mechanism Validation and Implementation Details

This appendix collects the additional diagnostics and clarifications requested during the review process.

N.1 Controlled Sweep over the Number of Updates per Rollout

To test the dominance reversal at intermediate off-policy degrees (beyond the two regimes of 2 and 16 updates per rollout), we run a controlled frozen-rollout experiment on Qwen2.5-3B with ORZ math. We fix the initial checkpoint, rollout tokens, response masks, behavior log-probabilities, advantages, update split, batch size, minibatch order, and the symmetric clipping rule ($\epsilon = 0.2$), and vary only the number $K \in \{2, 4, 8, 12, 16\}$ of optimizer updates applied to the same rollout (learning rate 5×10^{-5}). Tokens with $\pi_{\text{beh}} < 0.2$ are low-probability and those with $\pi_{\text{beh}} \geq 0.8$ are high-probability. Table 12 confirms that K tracks actual policy drift over all 71,633 valid tokens: both the mean absolute log-ratio and the token-weighted within-bin IS-ratio variance $\bar{\sigma}_\rho^2 = \sum_b (N_b/N)\, \widehat{\text{Var}}(\rho \mid b)$ (five fixed probability bins) increase monotonically with K.

Table 12: Measured policy drift on the frozen rollout as a function of the number of updates K.

K	2	4	8	12	16
$\mathbb{E}[\lvert \log \rho \rvert]$	0.358	0.522	0.620	0.752	1.137
$\bar{\sigma}_\rho^2$	7.02	41.04	100.92	153.34	293.08

We compute group-specific signed parameter gradients of the original seq-mean-token-mean clipped objective for probability group g, $\mathbf{g}_g(K) = \frac{1}{N_B} \sum_{i \in B} \frac{1}{T_i} \sum_t \mathbb{1}[p_{i,t} \in g]\, \nabla_\theta \ell_{i,t}^{\text{clip}}(\theta_K)$, and report $G_g(K) = \|\mathbf{g}_g(K)\|_2$. Signed token gradients are vector-summed before taking the norm, so cross-token direction and cancellation are preserved. As a module-local diagnostic we use the q_proj of the last transformer layer (4,196,352 parameters). We define the end-

point gap $D_{\log}(K) = \log\big(G_{[0,0.2)}(K)/G_{[0.8,1]}(K)\big)$ and the projected difference $\Delta_{\mathrm{proj}}(K) = \langle \mathbf{g}_{[0,0.2)} - \mathbf{g}_{[0.8,1]},\, \mathbf{g}_{\mathrm{all}} \rangle / \|\mathbf{g}_{\mathrm{all}}\|_2$, which accounts for alignment with the total module gradient.

Table 13: Group gradient norms ($\times 10^{-2}$) by behavior-probability bin, the endpoint gap $D_{\log}(K)$, and $\Delta_{\mathrm{proj}}(K)$ ($\times 10^{-3}$).

K	$G_{[0,0.2)}$	$G_{[0.2,0.4)}$	$G_{[0.4,0.6)}$	$G_{[0.6,0.8)}$	$G_{[0.8,1]}$	$D_{\log}(K)$	$\Delta_{\mathrm{proj}}(K)$
2	1.558	0.659	0.744	1.010	0.561	1.022	9.469
4	1.218	0.620	0.749	0.992	0.719	0.527	5.204
8	1.199	0.604	0.426	0.700	0.925	0.259	3.261
12	1.083	0.752	1.914	0.732	1.080	0.003	0.470
16	1.049	0.805	0.476	0.852	1.242	−0.169	−4.198

Both $D_{\log}(K)$ and $\Delta_{\mathrm{proj}}(K)$ decrease monotonically and change sign: the low-probability group produces the larger gradient at small K, and the high-probability group takes over by $K = 16$. The reversal is driven by the two endpoint groups, where the mechanism is sharpest; the middle bins form a noisier transition region, so we make no claim about a strict ordering across all five bins. Since the rollout, advantages, masks, bin membership, token counts, clipping rule, and update order are held fixed, this result is consistent with the dominance reversal predicted by Corollary 2.3. It is a single-rollout diagnostic on one module and one model, and we do not treat it as a general proof.

N.2 Empirical Advantage–Log-Ratio Association

The closed-form derivation uses a probability-conditional linear coupling $A = \mu(\pi)\, u + \eta$ with $u = \log \rho + \tau^2(\pi)/2$, in which both the slope $\mu(\pi)$ and the centering term depend on the behavior probability π; it is not a single global coupling. Moreover, the zero-residual condition ($\eta \equiv 0$) used for the closed form is an analytical simplification, not an empirical claim: η absorbs variation due to prompts, trajectories, token positions, and the broadcasting of one sequence-level advantage to many tokens.

To measure the association directly, we fix $K = 16$ and divide the update-split tokens into ten fixed-width behavior-probability bins. Within each bin b we use the binwise-centered log-ratio $\tilde{u}_{i,t} = \log \rho_{i,t} - \widehat{\mathbb{E}}[\log \rho \mid \pi \in b]$ and report the within-bin Pearson correlation between A and $\tilde{u}$ and the centered sign agreement $\Pr(A\tilde{u} > 0 \mid \pi \in b)$. Exact ties ($A = 0$ or $\tilde{u} = 0$) are excluded (none occur). Confidence intervals use 10,000 prompt-cluster bootstrap draws, with the bin mean re-estimated in every draw.

Table 14: Within-bin association between advantage and centered log-ratio at $K = 16$ (95% bootstrap CIs).

Behavior probability	Pearson [95% CI]	Centered sign agreement [95% CI]
$[0.0, 0.1)$	0.365 [0.273, 0.451]	0.590 [0.531, 0.663]
$[0.1, 0.2)$	0.303 [0.222, 0.402]	0.564 [0.512, 0.642]
$[0.2, 0.3)$	0.295 [0.234, 0.361]	0.573 [0.528, 0.629]
$[0.3, 0.4)$	0.265 [0.170, 0.352]	0.578 [0.529, 0.640]
$[0.4, 0.5)$	0.231 [0.156, 0.297]	0.558 [0.512, 0.609]
$[0.5, 0.6)$	0.208 [0.129, 0.291]	0.574 [0.514, 0.644]
$[0.6, 0.7)$	0.231 [0.176, 0.299]	0.575 [0.526, 0.626]
$[0.7, 0.8)$	0.241 [0.189, 0.306]	0.598 [0.543, 0.661]
$[0.8, 0.9)$	0.171 [0.104, 0.243]	0.551 [0.504, 0.609]
$[0.9, 1.0]$	0.058 [−0.035, 0.148]	0.567 [0.497, 0.650]

The within-bin correlations are positive but vary across bins (0.058 to 0.365) and are weakest for the highest-probability tokens, where the association is not statistically distinguishable from zero. Centered sign agreement is modest (0.551 to 0.598). These results support the qualitative content of the modeling assumption, a noisy but positive probability-conditional association, and confirm that exact zero-residual coupling is an idealization. The mechanism does not require exact coupling: the dominance reversal in Appendix N.1 is observed even though the measured association is only approximate.

N.3 Granularity of the Probability Bins

Frozen-batch diagnostic. Without retraining, we re-binned two frozen-rollout batches at $B \in \{2, 3, 5, 8, 10\}$ for each $K \in \{2, 4, 8, 12, 16\}$ and compared the resulting clip/no-clip decisions against the finest partition $B = 10$. Table 15 shows that the decision of a token changes for at most 0.593% of tokens between $B = 5$ and $B = 10$ and that no bin becomes sparse (the smallest bin still contains 954 tokens in the worst case over both batches and all K). For B from 2 to 10, the relative width of the 95% CI of $\widehat{\sigma}_b$ (10,000 prompt-cluster bootstrap draws) is at most 13.1% in every bin except the lowest-probability one, which is wider (57.6–67.6%) but does not grow with B, reflecting the intrinsic heavy tail of low-probability IS ratios rather than sparsity introduced by finer binning. This diagnostic is not an end-to-end RL ablation.

Table 15: Frozen-batch re-binning: smallest bin size and maximum clipping-decision mismatch relative to $B = 10$.

B	2	3	5	8	10
Minimum tokens per bin	8,858	3,375	2,000	1,227	954
Max. mismatch vs. $B = 10$	1.162%	1.162%	0.593%	0.152%	0%

End-to-end comparison on a matched budget. We extended the end-to-end bin-count comparison of Table 9 to $B \in \{2, 4, 5, 6, 10\}$ on Countdown with Qwen2.5-3B-Instruct. The $B = 2$ and $B = 10$ runs could not be trained to the full horizon, so every run in Table 16, including the DAPO baseline and the $B \in \{4, 5, 6\}$ runs, is restricted to the first 120 training steps and evaluated with the same metric as in the main text (best validation reward, mean@4). These numbers are therefore lower than the full-horizon results and are meaningful only as an internally consistent comparison. All adaptive configurations outperform the DAPO baseline in both regimes, with performance spans across B of 1.81 (N-OnP) and 1.23 (OffP) points and no monotonic degradation toward coarser or finer partitions. We do not claim that B is irrelevant or that $B = 5$ is uniquely optimal, and a full-horizon sweep is left to future work.

Table 16: Matched-budget (first 120 steps) bin-count comparison on Countdown (Qwen2.5-3B-Instruct), best validation reward (mean@4).

	DAPO	B=2	B=4	B=5	B=6	B=10
N-OnP	66.41	67.87	66.89	68.70	68.38	68.52
OffP	64.83	65.22	65.66	65.67	65.35	66.45

N.4 Implementation Details of ACPO

The bin-wise dispersion σ_b is a stop-gradient quantity. For each optimization mini-batch passed to the policy-loss computation, we pool all valid tokens of that mini-batch, compute the current IS ratios, estimate σ_b separately within each probability bin, detach the statistics, and construct the clipping bounds $\epsilon_b = \mathrm{clip}\big(\epsilon_{\mathrm{base}} + \alpha\,\mathrm{sg}[\sigma_b],\, \epsilon_{\min},\, \epsilon_{\max}\big)$, with $\epsilon_{\min} = 0$ and $\epsilon_{\max} = 3$ (Appendix I). Hence gradients propagate through neither the bin statistics nor the bounds. Under gradient accumulation, several such mini-batches contribute to one optimizer step; the bin statistics and clipping bounds are recomputed independently for each mini-batch and are not shared across an accumulated optimizer step or across the rollout-reuse cycle.

Scope of the ACPO rule. ACPO is an analysis-inspired implementation of probability-local dispersion calibration, not a rule derived from the closed-form gradient expectation. The analysis motivates why a single global boundary is mismatched to groups with different IS-ratio dispersion; the linear form $\epsilon_b = \epsilon_{\mathrm{base}} + \alpha\sigma_b$ (rather than, e.g., quantile-based widths or a target clip-rate controller), the use of the standard deviation, equal-width bins, and symmetric widening are simple design choices that we do not claim to be optimal. Uniform clipping rates across groups are not ACPO's objective: the clip rates serve only to diagnose whether a global threshold treats groups unfairly. Widening the boundary of a high-variance group can only lower its clipping frequency. ACPO retains clipping to cap very large IS ratios, which matters most when a rollout is reused for many updates, and we make

no claim that it outperforms clip-free methods such as GPG [Chu et al., 2025] or DisCO [Li et al., 2025]; a matched comparison is left to future work.

O Limitations

Magnitude vs. direction analyses are not on equal theoretical footing. The magnitude side of the analysis in Section 3 is grounded in the closed form for $\mathbb{E}[G \mid \pi]$ derived in Section 2.2, so its conclusions about which token populations dominate the effective update are model-level statements with explicit assumptions. The directional side, by contrast, is empirical: principal angles between gradient subspaces are measured directly on training runs, without an accompanying closed-form characterization of high-dimensional parameter-gradient direction. We treat the directional findings as diagnostic evidence consistent with the magnitude-side mechanism, rather than as theoretical claims; closed-form modeling of update direction is beyond the scope of this paper.

Theoretical results rest on idealizations. The closed forms in Proposition 2.2 and Corollary 2.3 are derived under modeling choices stated in Section 2.2 (log-normal IS ratio, linear advantage–log-ratio coupling, and the noiseless limit $\eta \equiv 0$). Robustness of the qualitative conclusions to the IS-ratio distribution is addressed in Appendix C.2.9 via a Gaussian re-derivation, and the finite-η situation is acknowledged in Appendix C.2.8; we do not develop a full finite-η closed form. Likewise, condition (C2) in Corollary 2.3 is verified for the representative (p, μ) in Appendix C.3 rather than for arbitrary distributions.

Empirical scope. Our experiments use the Qwen2.5 family (3B/3B-Instruct/7B) on math-reasoning and a small set of related tasks (ORZ, HiTab, Countdown). Cross-family generalization (e.g., to Llama-class or Mistral-class backbones) is not evaluated here, and we do not claim that ACPO's quantitative gains transfer unchanged across model families or task domains.

Scope: clipped, token-level IS. Our analysis and ACPO target the standard clipped, token-level IS setting (PPO/GRPO-style). The Section 3 results on token properties and advantage sign are policy-gradient statements that do not require clipping and would still apply to unclipped variants such as GPG [Chu et al., 2025]; in contrast, the IS-ratio variance, dominance reversal, and ACPO calibration in Sections 2 and 4 are specific to clipped objectives and do not directly carry over to unclipped methods. Likewise, our theory is for token-level IS; under sequence-level IS (e.g., GSPO [Zheng et al., 2025a]), the same dominance phenomenon would take a different form (which *sequences*, rather than which tokens, survive clipping), and a separate analysis would be needed. Sequence-level IS is also better suited to settings such as MoE training and rollout/training mismatch, which we do not address here.

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 (Section 1) state our three contributions—an analysis of how off-policy degree shapes IS-ratio dispersion and the effective gradient (Section 2), a component-wise study of the GRPO update (Section 3), and the ACPO algorithm (Section 4)—and these are matched by the corresponding theoretical and experimental sections.

   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: A dedicated Limitations section is provided in Appendix O, covering (i) the asymmetry between magnitude (closed-form) and direction (empirical) analyses, (ii) the modeling idealizations underlying our theoretical results, and (iii) the empirical scope of our experiments.

   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 (Lemma 2.1, Proposition 2.2, and Corollary 2.3) state their assumptions in the main text, and full derivations are given in Appendix C.1, Appendix C.2, and Appendix C.3.

   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: Section 5.1 describes the experimental setup, and Appendix I provides full training hyperparameters and per-method configurations. Code is released at `https://github.com/Control-derek/ACPO`.

   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: Code, training scripts, and configurations are released at `https://github.com/Control-derek/ACPO`; all datasets used (ORZ, HiTab, Countdown) are publicly available and cited in the paper.

   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: Appendix I reports the learning rate, batch sizes, rollout sizes, KL/entropy coefficients, loss-aggregation strategy, and all per-method hyperparameters; ACPO sensitivity analyses are provided in Appendix J.1.

   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: [Yes]

   Justification: Appendix M reports one-tailed Welch's t-tests (unequal variances) comparing ACPO against each baseline on the OOD math benchmarks, with the test statistic and assumptions stated explicitly.

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: Appendix L reports the hardware (one node with $4\times$NVIDIA A100 80GB GPUs) and approximate wall-clock time per training run (math: $\sim$14 days; HiTab: $\sim$7 days; Countdown: $\sim$2 days), and notes that preliminary and ablation runs consumed additional compute on the same hardware.

   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: We have reviewed the NeurIPS Code of Ethics; the work uses only publicly released models and datasets, and involves no human subjects or personal data (Appendix Ethics Statement).

   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: The Ethics Statement in the appendix discusses the positive impact (more reliable and predictable LLM reasoning) as well as the dual-use risks inherent to advancing LLM capabilities, and notes that our objective-reward training setup limits exposure to subjective bias.

    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: We do not release pre-trained models, scraped datasets, or generative-content artifacts; only the training-method code is released. The base models and datasets used are already publicly available under their original terms.

    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: All external assets used (Qwen2.5 base models, the ORZ, HiTab, and Countdown datasets, and the VeRL training framework) are cited in the references; their license and terms-of-use information is summarized in the README of the released code repository, and we use each asset within its stated terms.

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: The new asset introduced is the ACPO training code, released at `https://github.com/Control-derek/ACPO`. The repository includes a README documenting installation, configuration, training/evaluation scripts, and the hyperparameters needed to reproduce the reported results.

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 paper involves no crowdsourcing and no research with human subjects.

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 paper involves no human subjects, so IRB approval is not applicable.

    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: LLMs are the subject of study (we train them with RL), not a non-standard component of the methodology itself. Any use of LLMs in writing or editing the manuscript was limited to standard writing-aid purposes that, per the policy, do not require declaration.

    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.