TIC-GRPO: Provable and Efficient Optimization for Reinforcement Learning from Human Feedback

Lei Pang
Peking University
Beijing, China

Jun Luo
The Ohio State University
Columbus, USA

Ruinan Jin*
Mohamed bin Zayed University
of Artificial Intelligence
Abu Dhabi, UAE
jinruinan3@gmail.com

Abstract

Group Relative Policy Optimization (GRPO) trains language models without a value critic using rewards centered within sampled response groups. We study how importance weighting, clipping, and length normalization affect its stochastic updates and propose Trajectory-level Importance-Corrected GRPO (TIC-GRPO), combining an upper importance-ratio cap with a full-trajectory likelihood ratio. Under common maximum-length normalization, a trajectory change of measure gives a score second moment proportional to the response horizon. We propagate this estimate through the actual empirical baseline and all inner updates. With separately optimized constant steps, the sampling term in TIC's stationarity bound has a linear horizon coefficient, versus order three halves for token-level GRPO_2. Matching upper and lower bounds separate the capped methods' worst-case response-level variances by a factor T. The bounds track prompt count, response-group size, vocabulary size, reward scale, and score regularity; per-response-normalized GRPO also retains a length-covariance term. Experiments at two Qwen3 scales on four reasoning and coding benchmarks evaluate both mechanisms in sampled-token-normalized training.

1 Introduction

Reinforcement learning from human feedback (RLHF) aligns language-model outputs with task rewards and human preferences [Zhu et al., 2023, Bai et al., 2022]. Standard actor–critic Proximal Policy Optimization (PPO) [Schulman et al., 2017] uses a value critic; Group Relative Policy Optimization (GRPO) [Shao et al., 2024] constructs advantages from same-prompt response comparisons, enabling critic-free optimization of reasoning models.

Long responses sampled from an earlier policy make the update's stochastic structure important. Usual GRPO clipping can retain large upward importance ratios on negative-advantage samples. Capping controls this contribution, but clipped token scores generally have nonzero conditional means under the sampling policy. A full-trajectory likelihood ratio changes the entire response law, allowing the raw trajectory-score martingale to control the second moment.

We propose TIC-GRPO, which combines up-only clipping with trajectory-level importance correction and retains empirical group advantages. Its theoretical update uses the maximum response length T as a deterministic denominator. GRPO_2 provides a token-level comparator with the same upper cap and denominator. For both methods, the analysis retains the empirical baseline and the dependence of the reward on the entire response.

*Corresponding author.

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

The trajectory likelihood ratio transfers the TIC score second moment to the current policy, where the trajectory-score martingale gives a factor proportional to T. Direct Cauchy–Schwarz across the capped token scores gives a factor proportional to T^2 for GRPO_2. Our block ascent argument preserves this difference through separate step-size optimization and includes every inner iterate. With fixed reward scale, score regularity, cap, and vocabulary size, the resulting TIC bound is $\mathcal{O}(T/(N-1)) + \mathcal{O}(T\sqrt{(B^{-1}+|G|^{-1})/(N-1)})$. The corresponding token-level bound has $T^{3/2}$ in place of T in its square-root term. For original GRPO, normalization by each response's own length introduces an individual-length variance term without an explicit $1/|G|$ factor.

Contributions. We establish matching horizon orders for the response-level variance upper and lower bounds of GRPO_2 and TIC-GRPO, and an upper-tail instance with growing variance for original GRPO. We connect score moments of the actual empirical-baseline updates to averaged stationarity, accounting for clipping tails and all inner iterates under the specified conditional sampling protocol. We compare TIC-GRPO with DAPO-style GRPO [Yu et al., 2025] and GSPO [Zheng et al., 2025]. GSPO uses a length-normalized sequence importance ratio; TIC-GRPO uses the full trajectory ratio with an upper cap. Section 4.6 explains how these differences yield TIC-GRPO's moment-control advantage and compares their empirical update magnitudes. Experiments on Qwen3-1.7B and Qwen3-8B show improved scores on four mathematical-reasoning and coding benchmarks; ablations evaluate both algorithmic components.

Related work. Analyses of vanilla policy gradients establish sample-complexity guarantees under moment conditions [Yuan et al., 2022] and study convergence to stationary and locally optimal policies [Zhang et al., 2020]. PPO-Clip theory includes stationary-point convergence with potentially unbounded score functions [Jin et al., 2024a] and global convergence for deterministic, divergence-regularized algorithms with softmax policies [Dai et al., 2025b]. Heavy-tailed policy search relates stationarity and metastability to the policy's tail behavior [Bedi et al., 2024]. Related adaptive-optimization analyses study Adam [Kingma and Ba, 2014], establishing tight iteration-complexity bounds under smoothness and bounded variance [Wang et al., 2023] and convergence under moment conditions used for stochastic gradient descent [Jin et al., 2024b].

Research on language-model post-training examines alignment faking [Greenblatt et al., 2024], adaptive batch-wise sample scheduling for Direct Preference Optimization [Huang et al., 2025], and reward models that use evolving policy representations [Huang et al., 2026b]. Reasoning efficiency is addressed through sampling that exploits implicit stopping signals [Huang et al., 2026a] and hierarchical reinforcement learning that learns sampling temperatures from hidden states [Zhou et al., 2026]. Model efficiency is also studied through retraining-free expert pruning and neuron recombination in sparse mixture-of-experts language models [Zhou et al., 2025]. For group-relative methods, related analyses study advantage-estimation bias conditional on groups with mixed rewards [Yang et al., 2026], while work on diversity preservation in reinforcement learning with verifiable rewards investigates the divergence used for regularization [Li et al., 2025].

Applications include dual-clip PPO for complex control in multiplayer online battle arena games [Ye et al., 2020], PPO with self-generated rewards for Chinese spelling correction [Lin et al., 2026], and medical visual reasoning optimized with dual-stream GRPO [Zhu et al., 2025a]. Heterogeneous agents can share verified rollouts during training and act independently at inference [Zhang et al., 2026]. Multimodal modeling also develops dual implicit semantic and spatial memory for vision-language navigation [Zeng et al., 2025b], visual spatio-temporal chains of thought for autonomous driving [Zeng et al., 2025a], and language-guided prototype modulation for diabetic-retinopathy grading [Zhu et al., 2025b]. Mathematical-reasoning resources include the NuminaMath collection of competition problems and solutions [Li et al., 2024].

2 Preliminaries and the GRPO surrogate

2.1 Reinforcement learning setup

The policy parameter ranges over $\theta \in \mathbb{R}^l$. The prompt set is C; the prompt law p and terminal reward function are independent of this parameter. Responses use a finite vocabulary $\mathcal{V}$ and a positive integer horizon T. Let c be a prompt and let $a_t \in \mathcal{V}$ denote the t-th non-padding response token before termination. We write $s_t = (c, a_{\le t})$ for the prefix state conditioned on the prompt. For notational

convenience, variable-length responses are embedded in a fixed-horizon representation by appending a distinguished padding symbol PAD, which is masked out in all token-level sums:

$$s_T = (c, a_1, \ldots, a_t, \underbrace{\texttt{PAD}, \ldots, \texttt{PAD}}_{T-t \text{ positions}}), \qquad |s_T| = t.$$

Here $\mathcal{S}_T$ is the set of padded terminal responses and $|s_T|$ counts active response tokens. Padding after termination is deterministic, not a random action: its transition probability is one and its score is zero. Sums up to T are therefore masked over active tokens. Active transitions satisfy $\mathbb{P}_\theta(s_t \mid s_{t-1}) = \pi_\theta(a_t \mid s_{t-1})$, and terminal response probabilities factorize as

$$\mathbb{P}_\theta(s_T \mid c) = \prod_{t=1}^{|s_T|} \mathbb{P}_\theta(s_t \mid s_{t-1}) = \prod_{t=1}^{T} \mathbb{P}_\theta(s_t \mid s_{t-1}),$$

where the second product includes only unit-probability padding factors after termination. Rewards satisfy $|r(s_T)| \le R$. The objective and policy gradient are

$$J(\theta) := \mathbb{E}_{c\sim p,\, s_T\sim\mathbb{P}_\theta(\cdot|c)}[r(s_T)], \qquad J^* := \sup_\theta J(\theta) < \infty, \tag{1}$$

$$\nabla J(\theta) = \mathbb{E}_{c\sim p,\, s_T\sim\mathbb{P}_\theta(\cdot|c)}[\nabla \log \mathbb{P}_\theta(s_T \mid c)\, r(s_T)]\,. \tag{2}$$

Throughout, ∇ denotes gradients with respect to θ, unless stated otherwise.

2.2 Review of group relative policy optimization

This subsection fixes the GRPO-style token-level surrogate whose second moment will be analyzed later. Under a fixed old policy $\pi_{\theta_{\text{old}}}$, each inner batch contains B prompt occurrences drawn independently from p, conditional on the preceding iterates. Each occurrence of prompt c is assigned a group G_c of $|G|$ responses that are conditionally independent given the sampled prompts and the preceding iterates, with law $\mathbb{P}_{\theta_{\text{old}}}(\cdot \mid c)$. These prompt and response samples are fresh at each inner update. Let ξ_c be the empirical measure on $\mathcal{S}_T$: if s_T appears w times in G_c, then

$$\xi_c(s_T) = w/|G|, \tag{3}$$

where the group size is common across prompts. The group-relative advantage is

$$A_c(s_T) = r(s_T) - \mu_c, \qquad \mu_c = \sum_{s_T\in\mathcal{S}_T} \xi_c(s_T) r(s_T).$$

Following recent GRPO-style implementations [Liu et al., 2025, Chu et al., 2025, Dai et al., 2025a, DeepSeek-AI et al., 2025], we omit standard-deviation normalization of group-relative advantages: it can introduce additional bias [Liu et al., 2025, Chu et al., 2025] and is unnecessary for our variance comparison.

The response batch size is the total number of sampled responses per inner mini-batch: $B|G|$, from B prompts with $|G|$ responses each. For prompt-only functions, the empirical prompt measure $\zeta_k(c)$ is the prompt's occurrence count divided by B. Each occurrence, including repeated prompts, retains its own G_c, ξ_c, and μ_c. In occurrence sums, c denotes that occurrence's prompt; its group quantities are evaluated before averaging the B contributions. GRPO maximizes the group-relative surrogate

$$\begin{aligned}\mathrm{ClipMin}(s_t, \theta, \theta_{\text{old}}) := \min \Bigg\{ & \frac{\mathbb{P}_\theta(s_t \mid s_{t-1})}{\mathbb{P}_{\theta_{\text{old}}}(s_t \mid s_{t-1})} A_c(s_T), \\ & \mathrm{Clip}\left(\frac{\mathbb{P}_\theta(s_t \mid s_{t-1})}{\mathbb{P}_{\theta_{\text{old}}}(s_t \mid s_{t-1})}, \epsilon_{\text{low}}, \epsilon_{\text{high}}\right) A_c(s_T) \Bigg\},\end{aligned} \tag{4}$$

where $\mathrm{Clip}(x, \epsilon_{\text{low}}, \epsilon_{\text{high}}) := \min\{\max\{x, 1-\epsilon_{\text{low}}\}, 1+\epsilon_{\text{high}}\}$. The corresponding inner objective is

$$\mathcal{L}^{(k)}_{\mathrm{GRPO}}(\theta, \theta_{\text{old}}) := \frac{1}{B} \sum_{\text{prompt occurrences}} \sum_{s_T\in\mathcal{S}_T} \xi_c(s_T) \frac{1}{|s_T|} \sum_{t=1}^{T} \mathrm{ClipMin}(s_t, \theta, \theta_{\text{old}}). \tag{5}$$

The score-form GRPO direction is $\widehat{G}^{(k)}_{\mathrm{GRPO}} = \nabla \mathcal{L}^{(k)}_{\mathrm{GRPO}}$, with any subgradient between the one-sided values at clipping thresholds. The sampled rewards and empirical advantages are held fixed in this differentiation. The old policy is held fixed during the inner loop and refreshed after K updates. The sampled termination action retains its score when its probability depends on the parameter. Only the subsequent deterministic padding coordinates have zero score, so the sum to T is equivalent to summing over active tokens.

DAPO uses asymmetric clipping thresholds and omits the reference-policy KL penalty [Yu et al., 2025]. These choices are orthogonal to the variance mechanism: the surrogate remains token-level, with no uniform upper-tail cap on the negative-advantage branch. When $A_c(s_T) < 0$, Eq. (4) can select the unclipped ratio, admitting large upward ratios into the gradient. This incomplete upper-tail control is the first variance bottleneck addressed below. We use "GRPO" for the critic-free token-level mechanism shared by GRPO and DAPO in the theory.

2.3 Stochastic update structure

The empirical baseline, clipping multiplier, and response length are functions of the same sampled data. Their dependence is retained in the update analysis. TIC uses the trajectory-score martingale under its own policy law, together with a full-trajectory change of measure. GRPO$_2$ uses single-token conditional moments and Cauchy–Schwarz across the token positions.

3 Trajectory-level importance-corrected GRPO (TIC-GRPO)

GRPO combines token-level importance weighting with clipping. TIC-GRPO first caps upper-tail weights by up-only clipping, then replaces token-level weights by the full trajectory likelihood ratio. The intermediate comparator GRPO$_2$ retains token-level weighting with the same up-only cap and maximum-length denominator as TIC-GRPO.

Up-only clipping. For a token s_t on a trajectory s_T, define the one-step ratio $\rho_t(\theta, \theta_{\text{old}}) := \frac{\mathbb{P}_\theta(s_t|s_{t-1})}{\mathbb{P}_{\theta_{\text{old}}}(s_t|s_{t-1})}$. The standard GRPO clipping rule does not uniformly cap upper-tail ratios. Indeed, when $A_c(s_T) < 0$ and $\rho_t(\theta, \theta_{\text{old}}) > 1 + \epsilon_{\text{high}}$, the minimum in the GRPO surrogate can select the unclipped term $\rho_t(\theta, \theta_{\text{old}})A_c(s_T)$. We therefore use the up-only clipped ratio $\bar{\rho}_t(\theta, \theta_{\text{old}}) := \min\{\rho_t(\theta, \theta_{\text{old}}), 1 + \epsilon_{\text{high}}\}$. This clipping rule caps large upward ratios regardless of the sign of the advantage.

Intermediate comparator: GRPO$_2$. GRPO$_2$ uses up-only clipped token ratios and maximum-length normalization. Its k-th inner score-form update is

$$\widehat{G}^{(k)}_{\mathrm{GRPO}_2}(\theta, \theta_{\text{old}}) := \frac{1}{T}\frac{1}{B} \sum_{\text{prompt occurrences}} \sum_{s_T \in \mathcal{S}_T} \xi_c(s_T) \sum_{t=1}^{T} \bar{\rho}_t(\theta, \theta_{\text{old}}) A_c(s_T) \nabla \log \mathbb{P}_\theta(s_t \mid s_{t-1}), \tag{6}$$

where token sums are masked after termination and T is the maximum response length. The common cap and denominator isolate the effect of trajectory-level importance correction when comparing GRPO$_2$ with TIC-GRPO.

Token-score aggregation. The GRPO$_2$ feature sums capped token scores. Changing the next-token distribution bounds each token's conditional second moment under its old-policy prefix law. Cauchy–Schwarz across the T positions controls the complete feature, including all cross terms, yielding the moment estimate used in the convergence comparison.

Trajectory-level importance correction. TIC-GRPO replaces the token-level ratios by the trajectory likelihood ratio $\rho_{0:T}(\theta, \theta_{\text{old}}) := \frac{\mathbb{P}_\theta(s_T|c)}{\mathbb{P}_{\theta_{\text{old}}}(s_T|c)}$. The untruncated ratio $\rho_{0:T}$ is the Radon–Nikodym derivative for changing measure from old-policy trajectories to current-policy trajectories, and it pairs with the full trajectory score

$$G^\theta_T := \nabla \log \mathbb{P}_\theta(s_T \mid c) = \sum_{t=1}^{T} \nabla \log \mathbb{P}_\theta(s_t \mid s_{t-1}).$$

TIC-GRPO applies the same up-only clipping principle at the trajectory level: $\bar{\rho}_{0:T}(\theta, \theta_{\text{old}}) := \min\{\rho_{0:T}(\theta, \theta_{\text{old}}), 1 + \epsilon_{\text{high}}\}$. The clipped ratio is no longer an exact Radon–Nikodym derivative, but retains the needed likelihood structure: one copy of $\rho_{0:T}$ changes measure while the cap controls the remaining weight. The second moment thus reduces to a current-policy trajectory-score second moment, where martingale cancellation applies.

TIC-GRPO update. The k-th inner TIC-GRPO score-form direction is

$$\widehat{G}^{(k)}_{\text{TIC}}(\theta, \theta_{\text{old}}) := \frac{1}{T}\frac{1}{B} \sum_{\text{prompt occurrences}} \sum_{s_T \in \mathcal{S}_T} \xi_c(s_T)\bar{\rho}_{0:T}(\theta, \theta_{\text{old}})A_c(s_T)\nabla \log \mathbb{P}_\theta(s_T \mid c). \quad (7)$$

The parameter update is $\theta_{k+1} = \theta_k + \eta\, \widehat{G}^{(k)}_{\text{TIC}}(\theta_k, \theta_{\text{old}})$, with the old policy held fixed during the inner loop and refreshed after K updates, as in GRPO. Adam is widely used in practice [Kingma and Ba, 2014, Jin et al., 2026]; our analysis uses stochastic gradient ascent to isolate the RL update mechanisms.

Both capped updates use the maximum-length denominator T; responses may terminate earlier, and only active scores enter the numerator. Defined directly by (6) and (7), both score-form directions retain the capped multiplier at every ratio, including above the cap.

4 Variance analysis and martingale structure

We bound response-level variance cores above and below, then combine batch variances with mean-update stability. The upper bounds isolate clipping and trajectory-level importance weighting; binary constructions determine attainable horizon dependence.

Fix an arbitrary initial parameter $\theta_{1,0} \in \mathbb{R}^l$. We use the following outer–inner notation. The old policy is refreshed at the beginning of each subsequent outer iteration, $\theta_{n,0} = \theta_{n-1,K}, \qquad n \geq 2$, and $k = 0, \ldots, K-1$ indexes the inner updates performed under the fixed old policy $\theta_{n,0}$. For an algorithm $\mathsf{A} \in \{\text{GRPO}, \text{GRPO}_2, \text{TIC}\}$, we write the update in the common score-form notation $\theta_{n,k+1} = \theta_{n,k} + \eta\, \widehat{G}^{(n,k)}_{\mathsf{A}}$. Here $\widehat{G}^{(n,k)}_{\mathsf{A}}$ is the corresponding stochastic update direction: original GRPO uses its clipping-threshold subgradient, and the two capped methods use the score-form multipliers in Section 3. All mean and moment estimates retain the empirical baseline. The outer and within-block information fields are

$$\mathscr{F}_n = \sigma(\theta_{1,0}, \ldots, \theta_{n,0}), \qquad \mathscr{F}_{n,k} = \sigma(\mathscr{F}_n, \theta_{n,1}, \ldots, \theta_{n,k}).$$

We use the following assumptions throughout the variance and convergence analysis.

Assumption 4.1 (Lipschitz score in θ). There exists $L > 0$ such that for every active token transition $s_{t-1} \to s_t$ and all $\theta, \theta' \in \mathbb{R}^l$,

$$\left\|\nabla \log \mathbb{P}_\theta(s_t \mid s_{t-1}) - \nabla \log \mathbb{P}_{\theta'}(s_t \mid s_{t-1})\right\| \leq L\|\theta - \theta'\|.$$

The deterministic padding transitions have zero score and satisfy this condition trivially.

Assumption 4.2 (Bounded rewards and common support). There exists $R > 0$ such that $|r(s_T)| \leq R$ for all terminal states s_T. Moreover, all likelihood ratios and logarithms appearing in the analysis are evaluated on a common positive support: for every active transition on this support and every $\theta \in \mathbb{R}^l$,

$$\mathbb{P}_\theta(s_t \mid s_{t-1}) > 0, \qquad \mathbb{P}_{\theta_{n,0}}(s_t \mid s_{t-1}) > 0.$$

Transitions after termination are deterministic and have zero score; a sampled termination action is included in the likelihood. All response lengths satisfy $1 \leq |s_T| \leq T$.

Related policy-gradient analyses impose regularity on log policies and their scores [Zhang et al., 2023, Xiong et al., 2021, Ding et al., 2021, Suk and Duan, 2025]. Here the score itself need not be uniformly bounded. Common support ensures that the ratios and logarithms are well defined.

The theoretical sampling protocol is specified as follows. Conditional on $\mathscr{F}_{n,k}$, the $B \geq 1$ prompt occurrences in an inner batch are independent draws from p, each contributing weight $1/B$. Thus the empirical prompt measure satisfies $\mathbb{E}[\zeta_k(c) \mid \mathscr{F}_{n,k}] = p(c)$. Given the sampled prompts and $\mathscr{F}_{n,k}$, the $|G|$ responses within each occurrence are independent draws from $\mathbb{P}_{\theta_{n,0}}(\cdot \mid c)$. The response

groups are freshly sampled at each inner update. Dependence between different response groups is allowed. Each empirical baseline and sampled response length belong to the same group. This conditional sampling protocol is used throughout the stationarity analysis. We use $\mathbb{E}_\theta$ for the joint expectation over $c \sim p$ and $s_T \sim \mathbb{P}_\theta(\cdot \mid c)$; a displayed conditioning on c retains only the response expectation. The prompt reward mean is $\mu_{\theta,c} = \mathbb{E}_\theta[r(s_T) \mid c]$. The cap satisfies $\epsilon_{\text{high}} \geq 0$; the GRPO drift bound with reciprocal clipping margins is stated for positive margins.

4.1 Trajectory scores and martingale cancellation

Fix a prompt c and a policy parameter θ. Define the one-step score and the full trajectory score as

$$g_t^\theta := \nabla \log \mathbb{P}_\theta(s_t \mid s_{t-1}), \qquad G_T^\theta := \nabla \log \mathbb{P}_\theta(s_T \mid c) = \sum_{t=1}^{T} g_t^\theta,$$

where the sum is masked after termination. Under the current-policy trajectory law $\mathbb{P}_\theta(\cdot \mid c)$, the one-step score terms form a martingale-difference sequence: $\mathbb{E}_\theta[g_t^\theta \mid s_{t-1}] = 0$. Consequently, the cross terms vanish in the squared trajectory score.

Property 4.1 (Trajectory-score second moment)**.** *For any prompt c and any θ,*

$$\mathbb{E}_{s_T \sim \mathbb{P}_\theta(\cdot|c)}\left[\|G_T^\theta\|^2\right] = \mathbb{E}_{s_T \sim \mathbb{P}_\theta(\cdot|c)}\left[\sum_{t=1}^{T} \|g_t^\theta\|^2\right].$$

For TIC, the full-trajectory likelihood ratio changes the sampling law before this identity is applied. For GRPO$_2$, the update core is bounded by Cauchy–Schwarz across token positions and a separate next-token change of measure at each prefix.

4.2 Variance cores for GRPO, GRPO$_2$, and TIC-GRPO

Let

$$\rho_t := \frac{\mathbb{P}_{\theta_{n,k}}(s_t \mid s_{t-1})}{\mathbb{P}_{\theta_{n,0}}(s_t \mid s_{t-1})}, \qquad \rho_{0:T} := \frac{\mathbb{P}_{\theta_{n,k}}(s_T \mid c)}{\mathbb{P}_{\theta_{n,0}}(s_T \mid c)}, \qquad c_+ := 1 + \epsilon_{\text{high}}.$$

We write the up-only clipped token and trajectory ratios as $\rho_t^+ := \min\{\rho_t, c_+\}$, $\rho_{0:T}^+ := \min\{\rho_{0:T}, c_+\}$. For original GRPO, let m_t^{GRPO} denote the multiplier that appears in the score-form core of the original clipped token-level update. Away from the clipping thresholds, this multiplier is either zero or the active token ratio selected by the GRPO clipping rule; at a threshold, we take any subgradient in the closed interval between the two one-sided limits. Because the negative-advantage branch can select an unclipped upper-tail ratio, the multiplier is controlled on the sampled old-policy support only through

$$|m_t^{\text{GRPO}}| \leq \mathcal{M}_n, \qquad \mathcal{M}_n := \max_{s_T \in \mathcal{S}_T^{(n)}} \max_{1 \leq t \leq |s_T|} \frac{1}{\mathbb{P}_{\theta_{n,0}}(s_t \mid s_{t-1})}. \tag{8}$$

Here $\mathcal{S}_T^{(n)}$ contains the responses sampled during outer iteration n, including their prompt labels. The factor $\mathcal{M}_n$ is a random maximum over their active transitions and controls the upper-tail contribution in GRPO.

We compare the following stochastic update cores:

$$\begin{gathered} Z_{\text{GRPO}} := \frac{A_c(s_T)}{|s_T|} \sum_{t=1}^{|s_T|} m_t^{\text{GRPO}} g_t^{\theta_{n,k}}, \qquad Z_{\text{GRPO}_2} := \frac{A_c(s_T)}{T} \sum_{t=1}^{T} \rho_t^+ g_t^{\theta_{n,k}}, \\ Z_{\text{TIC}} := \frac{A_c(s_T)}{T} \rho_{0:T}^+ G_T^{\theta_{n,k}}. \end{gathered} \tag{9}$$

Each core is a single response's contribution to the batch update direction. Empirical averaging over prompts and responses gives $\widehat{G}_{\mathsf{A}}^{(n,k)}$. The direct score-form update therefore satisfies $\|\theta_{n,k+1} - \theta_{n,k}\|_2^2 = \eta^2 \|\widehat{G}_{\mathsf{A}}^{(n,k)}\|_2^2$, the squared Euclidean distance between consecutive parameter iterates. By

Jensen's inequality, each conditional core second-moment upper bound, multiplied by η^2, also bounds the corresponding conditional expected squared distance.

These three cores isolate the variance mechanisms of interest: original GRPO may retain uncontrolled upper-tail multipliers, GRPO_2 caps upper-tail ratios but remains token-level, and TIC-GRPO applies the same upper-tail cap at the trajectory level.

4.3 Upper bounds for the three variance cores

Proposition 4.1 (Core update moments, informal)**.** *Under Assumptions 4.1–4.2 and the stated sampling protocol, the conditional second moments of Z_{GRPO}, Z_{GRPO_2}, and Z_{TIC}, given $\mathscr{F}_{n,k}$, are respectively*

$$\mathcal{O}\big(R^2 L\,\mathbb{E}[\mathcal{M}_n^2 \mid \mathscr{F}_{n,k}]\big)\,, \qquad \mathcal{O}(R^2 L c_+ \log|\mathcal{V}|), \qquad \mathcal{O}\left(\frac{R^2 L c_+ \log|\mathcal{V}|}{T}\right).$$

The constants are universal; the first bound is interpreted in the extended real sense. The distinguished response uses the empirical prompt and response weights, and the same bounds apply to the empirical average update by Jensen's inequality.

Proposition E.1 in Appendix E gives the full statement with explicit constants and its proof.

Each displayed second-moment bound also bounds the corresponding conditional variance, since subtracting the conditional mean reduces the second moment. With R, L, c_+ and $|\mathcal{V}|$ fixed, the capped response-level cores therefore have variance upper bounds of order 1 for GRPO_2 and $1/T$ for TIC-GRPO. For original GRPO, the dependence on the old-policy inverse probabilities remains explicit through $\mathcal{M}_n$.

For TIC-GRPO the key inequality is $(\rho^+_{0:T})^2 \le c_+ \rho_{0:T}$. The trajectory ratio changes the law to the current policy, under which Property 4.1 applies. These bounds describe the two importance-weighting mechanisms under a common denominator. The convergence analysis additionally separates the prompt and response contributions to the batch variance and controls the complete mean update within each outer block.

4.4 Lower bounds and tightness of the variance scaling

The horizon dependence of the capped response-level bounds is attained by binary policies with bounded rewards. The constructions retain the empirical group baseline and the normalizations in the three expressions of Eq. (9).

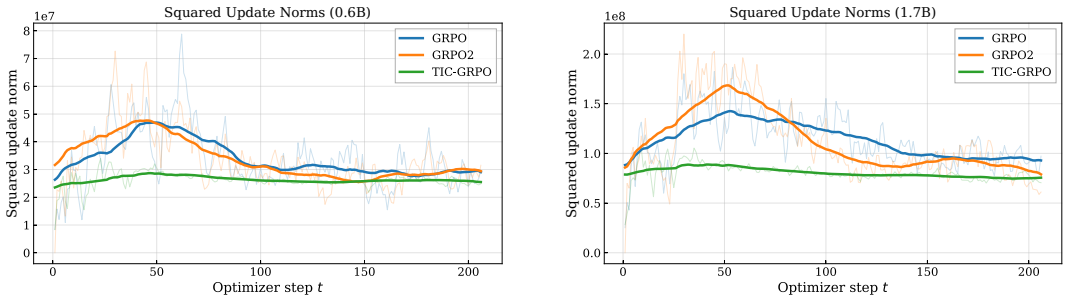
Proposition 4.2 (Response-level variance lower bounds, informal)**.** *For every $R, L > 0$, $c_+ \ge 1$, $T \ge 2$, and $|G| \ge 2$, separate one-dimensional binary policy instances satisfying Assumptions 4.1–4.2, with one prompt, independent responses, and deterministic length T, attain conditional response-core variances*

$$\mathrm{GRPO}_2: \quad \Omega\left(R^2 L\left(1 - \frac{1}{|G|}\right)\right), \qquad \mathrm{TIC}: \quad \Omega\left(\frac{R^2 L}{T}\left(1 - \frac{1}{|G|}\right)\right).$$

Original GRPO has a binary instance with variance $\Omega(R^2 L T)$ when $|G| = 2$, $T \ge 64$, and $T^2 > 2c_+$. The constants are universal. All cores are evaluated at fixed old/current parameter pairs before sampling. The old-policy success probabilities are T^{-2} for original GRPO and $1/(4c_+)$ for GRPO_2.

Proposition F.1 in Appendix F gives the explicit lower bounds and their complete constructions.

Together, Propositions 4.1 and 4.2 give matching horizon orders 1 and $1/T$ for the worst-case response-level variances of GRPO_2 and TIC-GRPO, respectively, with fixed $|G| \ge 2$ and a binary vocabulary. Equivalently, the second moments of their unnormalized numerators have orders T^2 and T. Original GRPO admits a variance of order at least T as its old-policy upper-tail ratios grow, and the same construction gives an unnormalized second moment of order at least T^3. These statements concern the individual response cores; the empirical batch averages and the iterates generated by the algorithms are analyzed in Section 5.



(a) Qwen3-0.6B. (b) Qwen3-1.7B.

Figure 1: Squared Euclidean distances between consecutive parameter iterates, $\|\theta_{k+1} - \theta_k\|_2^2$. TIC-GRPO has the lowest and most stable values in these runs.

4.5 Empirical update magnitudes

To examine update magnitudes in practical LLM fine-tuning, we track $\|\theta_{k+1} - \theta_k\|_2^2$, the squared Euclidean distance between consecutive parameter iterates, on Qwen3-0.6B and Qwen3-1.7B. For a fixed learning rate and a direct stochastic-gradient update, its expectation equals the batch update direction's second moment multiplied by the squared learning rate. In the experiments, this distance is computed from the actual parameter iterates produced by the optimizer. All compared methods use the same data and optimizer settings from Appendix A.1.

The curves measure realized update sizes under the experimental optimizer and provide an empirical comparison of the three training rules.

4.6 Comparison with GSPO

GSPO applies PPO-style minimum clipping to the root-normalized sequence ratio $\rho_{0:T}^{1/|s_T|}$. Negative advantages can retain arbitrarily large root ratios above the upper clipping threshold. TIC-GRPO instead satisfies $(\rho_{0:T}^{+})^2 \le c_+ \rho_{0:T}$ regardless of the advantage's sign: the cap controls one weight factor, and the exact trajectory ratio transfers the remaining score moment to the current policy, where Property 4.1 applies. GSPO's root ratio generally differs from this density ratio. Even if the root ratio were capped, applying the same rescaling-and-transfer argument would leave $\rho_{0:T}^{1/|s_T|-1}$ inside the current-policy expectation, which is unbounded as $\rho_{0:T} \to 0$ for $|s_T| > 1$. TIC combines the two properties supporting the uniform moment reduction in Proposition 4.1.

Matched runs with the same data, optimizer, and training settings over steps 1–200 on Qwen3-0.6B and Qwen3-1.7B give TIC-GRPO about 21% and 26% lower temporal means of the realized squared update norm than GSPO, respectively. The corresponding standard deviations across training steps are about 69% and 76% lower, indicating smaller and more stable realized updates. Appendix A.4 reports the protocol and interval-level results in Table 3.

5 Convergence guarantees

Stationarity is averaged over all pre-update parameters, including the initial parameter and excluding the final post-update parameter. Both capped methods use the common maximum response length T as their deterministic denominator. Expectations are along the run of the algorithm under consideration.

5.1 Convergence with independent prompt sampling

Independent prompt draws and empirical group covariance yield separate B and $|G|$ contributions, with explicit coefficients from the directional score second moment and its trace, respectively. Dependence between response groups is permitted, so response batch size $B|G|$ cannot replace both sample counts in the variance bound.

Theorem 5.1 (TIC-GRPO and GRPO$_2$ convergence, informal)**.** *Under Assumptions 4.1–4.2 and the conditional sampling protocol, let $N \ge 2$, $K \ge 1$, $B \ge 1$, $|G| \ge 2$, and $|\mathcal{V}| \ge 2$. With separately*

selected admissible constant steps and fixed $R, L, c_+, |\mathcal{V}|$, the average squared gradient norm

$$\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla J(\theta_{n,k})\right\|^2$$

is bounded by the following respective orders:

$$\text{TIC}: \quad \mathcal{O}\left(\frac{T}{N-1}\right)+\mathcal{O}\left(T\sqrt{\frac{B^{-1}+|G|^{-1}}{N-1}}\right), \tag{10}$$

$$\text{GRPO}_2: \quad \mathcal{O}\left(\frac{T}{N-1}\right)+\mathcal{O}\left(T^{3/2}\sqrt{\frac{B^{-1}+|G|^{-1}}{N-1}}\right). \tag{11}$$

The constants are independent of T, N, K, B, and $|G|$. Expectations follow each algorithm's own run, and both guarantees include a zero initial gap.

Theorem C.1 in Appendix C gives the finite-step inequalities, full parameter dependence, admissible step-size interval, and separate step choices.

After separate step-size optimization, the sampling-term horizon coefficients are T for TIC and $T^{3/2}$ for the token-level guarantee, a $\sqrt{T}$ improvement; the common finite-iteration term is linear in T. This compares upper guarantees, not performance lower bounds. It holds for the same $R, L, c_+, K, B, |G|, |\mathcal{V}|$ and an initial gap bounded independently of T, with $J^* - J(\theta_{1,0}) \le 2R$. The explicit bounds track all dependencies; any variation of batch or vocabulary size with T must also be included.

The square-root term determines the horizon comparison when it dominates the common finite-iteration term. Both methods have the same update at $K = 1$, because every update is on policy; the two displayed bounds therefore represent different upper guarantees for that case. They also coincide at $T = 1$. For $|G| = 1$, all empirical advantages vanish and the universal bound R^2LT applies. The cases $R = 0$, $L = 0$, or $|\mathcal{V}| = 1$ have zero objective gradient. The step-size interval in Theorem C.1 is sufficient for the block ascent argument.

5.2 Per-response normalization

For original GRPO, the joint old-policy response-length variance is

$$\sigma^2_{\theta_{n,0}} := \mathrm{Var}_{c\sim p,\, s_T\sim\mathbb{P}_{\theta_{n,0}}(\cdot|c)}(|s_T|). \tag{12}$$

Its per-response denominator introduces a length-covariance contribution to the population update.

Theorem 5.2 (GRPO convergence, informal)**.** *Under the same sampling protocol and regularity assumptions, let $N \ge 2$, $K \ge 1$, $|G| \ge 2$, $\eta > 0$, and $\epsilon_{low}, \epsilon_{high} > 0$. For $K \ge 2$, original GRPO satisfies*

$$\begin{aligned}\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla J(\theta_{n,k})\right\|^2 &\le \mathcal{O}\left(\frac{T}{\eta(N-1)}\right)+\mathcal{O}\left(\frac{\eta T^2}{N-1}\sum_{n=1}^{N-1}\mathbb{E}[\mathcal{M}_n^2]\right)\\ &+\mathcal{O}\left(\eta T^{5/2}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{E}[\mathcal{M}_n^2]}\right)+\mathcal{O}\left(\frac{T}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right]\right).\end{aligned} \tag{13}$$

Constants depend only on $R, L, K, \epsilon_{low}, \epsilon_{high}, |\mathcal{V}|$ and $J^* - J(\theta_{1,0})$. For $K = 1$, suitable constant steps, including when the initial gap is zero, give

$$\frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\left\|\nabla J(\theta_{n,0})\right\|^2 \le \mathcal{O}\left(\frac{T^{3/2}}{\sqrt{N-1}}\right)+\mathcal{O}\left(\frac{T}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right]\right). \tag{14}$$

Theorem D.1 in Appendix D gives the explicit inequalities and step choices. For a positive initial gap, the $K = 1$ choice minimizes the two explicit optimization terms; the length term remains evaluated along the resulting run.

The individual-length term from per-response normalization is at most a constant times $T(\sqrt{T}-1)^2$, without a B^{-1} or $|G|^{-1}$ factor. For general K, the inverse-probability moments in (13) describe the original clipping rule's upper tail; they remain inside expectations along the run at the selected step size. For $K = 1$, all updates are on policy, giving (14) directly from the score moment. All bounds can be capped by R^2LT, including when a clipping margin is zero and the ratio-one subgradient is unspecified.

6 Experiments

We evaluate TIC-GRPO on Qwen3-1.7B and Qwen3-8B [Yang et al., 2025] against GSPO [Zheng et al., 2025] and a DAPO implementation of the critic-free group-relative GRPO baseline [Yu et al., 2025]. Table 1 shows that TIC-GRPO is best on all four benchmarks; Appendix A.1–A.3 reports setup details, learning curves, and evaluations of the component variants.

Table 1: Main results. Avg@1 uses greedy decoding and Avg@32 uses 32 samples; parentheses are gains over GRPO.

Method	**AIME24** (Avg@32)	**AIME25** (Avg@32)	**MATH500** (Avg@1)	**Live-CodeBench** (Avg@1)
		Qwen3-1.7B		
GRPO	9.17	5.31	66.6	10.67
GSPO	10.31 (+1.14)	6.24 (+0.93)	69.0 (+2.4)	11.23 (+0.56)
TIC-GRPO	**11.77 (+2.60)**	**6.98 (+1.67)**	**69.8 (+3.2)**	**12.11 (+1.44)**
		Qwen3-8B		
GRPO	31.35	22.9	88.6	19.88
GSPO	30.21 (-1.14)	22.5 (-0.4)	88.4 (-0.2)	20.00 (+0.12)
TIC-GRPO	**33.34 (+1.99)**	**24.12 (+1.22)**	**90.0 (+1.4)**	**21.00 (+1.12)**

7 Conclusion

TIC-GRPO combines upper-tail clipping with trajectory-level importance correction in a critic-free group-relative update. Under common maximum-length normalization, matching upper and lower bounds give worst-case response-level variance orders 1 for GRPO_2 and $1/T$ for TIC-GRPO. A full-trajectory change of measure preserves the score-martingale second moment through the empirical-baseline and block-convergence analysis. With separately optimized step sizes, the sampling term in TIC's stationarity bound has a linear horizon coefficient, improving by $\sqrt{T}$ over the displayed token-level GRPO_2 guarantee. The bounds separate prompt and within-group sampling. Experiments and ablations evaluate both mechanisms in sampled-token-normalized training on reasoning and coding benchmarks.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding or in-kind support for this work. The authors declare no competing financial interests.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Amrit Singh Bedi, Anjaly Parayil, Junyu Zhang, Mengdi Wang, and Alec Koppel. On the sample complexity and metastability of heavy-tailed policy search in continuous control. *Journal of Machine Learning Research*, 25(39):1–58, 2024.

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.

Muzhi Dai, Chenxu Yang, and Qingyi Si. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models. *arXiv preprint arXiv:2505.07686*, 2025a.

Qiming Dai, Yin Liu, Junyu Zhang, and Zaiwen Wen. Non-Asymptotic Global Convergence of PPO-Clip. *arXiv preprint arXiv:2512.16565*, 2025b.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, et al. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. *arXiv preprint arXiv:2512.02556*, 2025.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the Global Optimum Convergence of Momentum-based Policy Gradient. *arXiv preprint arXiv:2110.10116*, 2021.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

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.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. Adaptive Batch-Wise Sample Scheduling for Direct Preference Optimization. *arXiv preprint arXiv:2506.17252*, 2025.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, Fuzhen Zhuang, Jianxin Li, Deqing Wang, and Yikun Ban. Does Your Reasoning Model Implicitly Know When to Stop Thinking? *arXiv preprint arXiv:2602.08354*, 2026a.

Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuefeng Xiao, Hongyan Xie, Li Huaqiu, Songshi Liang, Zhongxiang Dai, Fuzhen Zhuang, Jianxin Li, Yikun Ban, and Deqing Wang. Real-Time Aligned Reward Model beyond Semantics. *arXiv preprint arXiv:2601.22664*, 2026b.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code. *arXiv preprint arXiv:2403.07974*, 2024.

Ruinan Jin, Shuai Li, and Baoxiang Wang. On Stationary Point Convergence of PPO-Clip. In *International Conference on Learning Representations*, 2024a.

Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A Comprehensive Framework for Analyzing the Convergence of Adam: Bridging the Gap with SGD. *arXiv preprint arXiv:2410.04458*, 2024b.

Ruinan Jin, Difei Cheng, Ling Chen, Jun Luo, Hao Zhou, and Youzhi Zhang. Adam under Generalized Smoothness with Second-Moment-Type Stochastic Gradients. *arXiv preprint arXiv:2609.37787*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath: The Largest Public Dataset in AI4Maths with 860K Pairs of Competition Math Problems and Solutions. Technical report, Project Numina, 2024.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Klein Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The Choice of Divergence: A Neglected Key to Mitigating Diversity Collapse in Reinforcement Learning with Verifiable Reward. *arXiv preprint arXiv:2509.07430*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *arXiv preprint arXiv:2305.20050*, 2023.

Zhiming Lin, Kai Zhao, Sophie Zhang, Peilai Yu, and Canran Xiao. CEC-Zero: Zero-supervision character error correction with self-generated rewards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23612–23620, 2026. doi: 10.1609/aaai.v40i28.39534.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Mathematical Association of America. American Invitational Mathematics Examination (AIME). MAA Invitational Competitions, n.d. Accessed September 20, 2026.

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, Y. K. Li, Y. Wu, and Daya Guo. 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.

Joe Suk and Yaqi Duan. On the optimization dynamics of RLVR: Gradient gap and step size thresholds. *arXiv preprint arXiv:2510.08539*, 2025.

Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *Advances in Neural Information Processing Systems*, volume 36, pages 39006–39032, 2023.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic Convergence of Adam-type Reinforcement Learning Algorithms under Markovian Sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10460–10468, 2021. doi: 10.1609/aaai.v35i12.17252.

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

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your Group-Relative Advantage Is Biased. *arXiv preprint arXiv:2601.08521*, 2026.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering Complex Control in MOBA Games with Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 6672–6679, 2020. doi: 10.1609/aaai.v34i04.6144.

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.

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3332–3380. PMLR, 2022.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, Xing Wei, and Ning Guo. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv preprint arXiv:2505.17685*, 2025a.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, Xing Wei, and Ning Guo. JanusVLN: Decoupling Semantics and Spatiality with Dual Implicit Memory for Vision-Language Navigation. *arXiv preprint arXiv:2509.22548*, 2025b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020. doi: 10.1137/19M1288012.

Shenao Zhang, Boyi Liu, Zhaoran Wang, and Tuo Zhao. Model-based reparameterization policy gradient methods: Theory and practical algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pages 68391–68419, 2023.

Zhixia Zhang, Zixuan Huang, Gongxun Li, Huaiyang Wang, Chengyi Yuan, Xin Xia, Deqing Wang, Fuzhen Zhuang, Shuai Ma, Ning Ding, Yaodong Yang, Jianxin Li, and Yikun Ban. Heterogeneous Agent Collaborative Reinforcement Learning. *arXiv preprint arXiv:2603.02604*, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Yixiao Zhou, Ziyu Zhao, Dongzhou Cheng, Zhiliang Wu, Jie Gui, Yi Yang, Fei Wu, Yu Cheng, and Hehe Fan. Dropping Experts, Recombining Neurons: Retraining-Free Pruning for Sparse Mixture-of-Experts LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15169–15186, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.820.

Yixiao Zhou, Yang Li, Dongzhou Cheng, Hehe Fan, and Yu Cheng. Look Inward to Explore Outward: Learning Temperature Policy from LLM Internal States via Hierarchical RL. *arXiv preprint arXiv:2602.13035*, 2026.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 2023.

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. MedEyes: Learning Dynamic Visual Focus for Medical Progressive Diagnosis. *arXiv preprint arXiv:2511.22018*, 2025a.

Chunzheng Zhu, Yangfang Lin, Jialin Shao, Jianxin Lin, and Yijun Wang. Pathology-Aware Prototype Evolution via LLM-Driven Semantic Disambiguation for Multicenter Diabetic Retinopathy Diagnosis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9196–9205, 2025b. doi: 10.1145/3746027.3754562.

A Experiments

A.1 Experimental setup

Models and Training Data. We conduct all experiments using the Qwen3 [Yang et al., 2025] family of models, including Qwen3-0.6B, Qwen3-1.7B, and Qwen3-8B. The training data consist of a hybrid dataset that combines the full DAPO-Math-17K [Yu et al., 2025] corpus with a subset of the AIME benchmark (1983–2022), resulting in several hundred additional high-difficulty mathematical reasoning problems. All RL training is conducted within the VeRL [Sheng et al., 2024] framework, and all models are trained for a single epoch, ensuring that each prompt is used exactly once during training. We use a DAPO-style implementation as the experimental GRPO baseline and implement the proposed variants within the same training framework. DAPO combines the critic-free group-relative update with asymmetric clipping and a token-level loss [Yu et al., 2025]. The experimental implementation divides the capped update numerator by the sampled mean number of active response tokens. The theoretical results concern the maximum-length-normalized updates and fresh conditional sampling; the experiments evaluate the clipping and trajectory-correction mechanisms with sampled-token normalization and trajectory reuse.

Evaluation. We evaluate on three mathematical reasoning benchmarks and one coding benchmark: MATH500, the 500-problem subset of MATH introduced by Lightman et al. [2023] from the dataset of Hendrycks et al. [2021]; the 2024 and 2025 American Invitational Mathematics Examinations (AIME24 and AIME25) [Mathematical Association of America, n.d.]; and LiveCodeBench [Jain et al., 2024]. We report Avg@1 on MATH500 and LiveCodeBench. For AIME24 and AIME25, which each contain 30 problems, we report Avg@32.

- **Avg@1:** Standard accuracy computed using greedy decoding, measuring deterministic single-response performance.
- **Avg@32:** Average accuracy over 32 sampled responses per problem, obtained with a temperature of 1.0 and top-p of 1.0. This metric estimates mean accuracy under the specified sampling distribution.

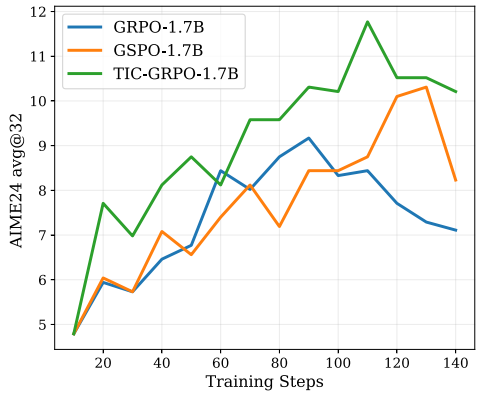
Training Configuration. We adopt a total batch size of 128 with a mini-batch size of 32. Each sampled trajectory is reused for four gradient updates before the old policy is refreshed. This configuration enables efficient reuse of collected samples while maintaining stable optimization. Training of Qwen3-1.7B is performed on a single H200 GPU for approximately 24 hours, whereas Qwen3-8B is trained on two H200 nodes for 48 hours. For the larger model, gradient accumulation is applied to ensure that the effective global batch size remains consistent with that of Qwen3-1.7B. Furthermore, to eliminate confounding factors introduced by stochastic data processing, we disable both Dynamic Sampling and Soft Overlong Punishment in all experiments. Unless otherwise specified, all reported results follow this unified experimental configuration.

A.2 Primary empirical results of TIC-GRPO

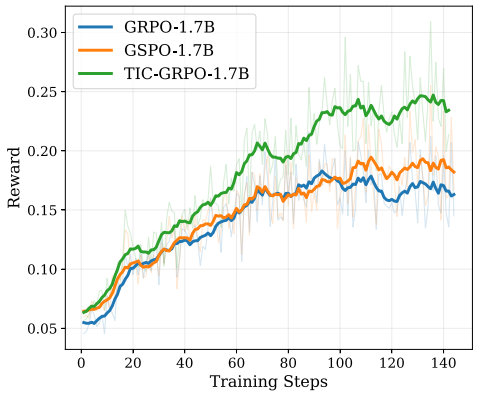
As concluded in Sec. 6, our proposed TIC-GRPO consistently outperforms recent advanced GRPO-style baselines [Yu et al., 2025, Zheng et al., 2025] across different model scales. In this section, we further present the training reward and accuracy curves corresponding to the results in Table 1, together with additional experimental evidence that provides a more comprehensive evaluation of TIC-GRPO.

Figure 2 summarizes the primary empirical results on both Qwen3-1.7B and Qwen3-8B models, where each row corresponds to a different model size. For each model, we report the AIME24 Avg@32 accuracy curve (left) alongside the corresponding training reward curve (right), showing the evolution of evaluation accuracy and training reward.

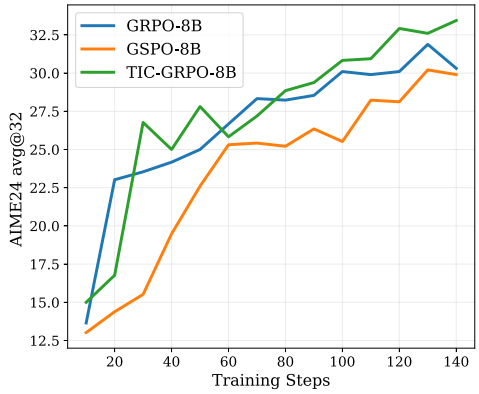
TIC-GRPO attains the largest reported AIME24 scores in Table 1 at both model scales. Figure 2 shows its evaluation accuracy and training reward over the course of optimization. The accuracy curves cross during training, and the relative ordering of training rewards can differ from that of evaluation accuracy.



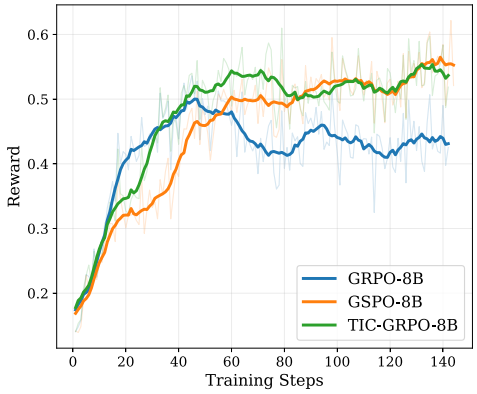
(a) AIME24 Avg@32 accuracy on Qwen3-1.7B.



(b) Training reward curves on Qwen3-1.7B.



(c) AIME24 Avg@32 accuracy on Qwen3-8B.



(d) Training reward curves on Qwen3-8B.

Figure 2: Training dynamics of different GRPO variants on Qwen3-1.7B and Qwen3-8B models. Left panels show the AIME24 Avg@32 accuracy curves, while right panels report the corresponding training reward curves. Each panel compares the three methods at the recorded training steps.

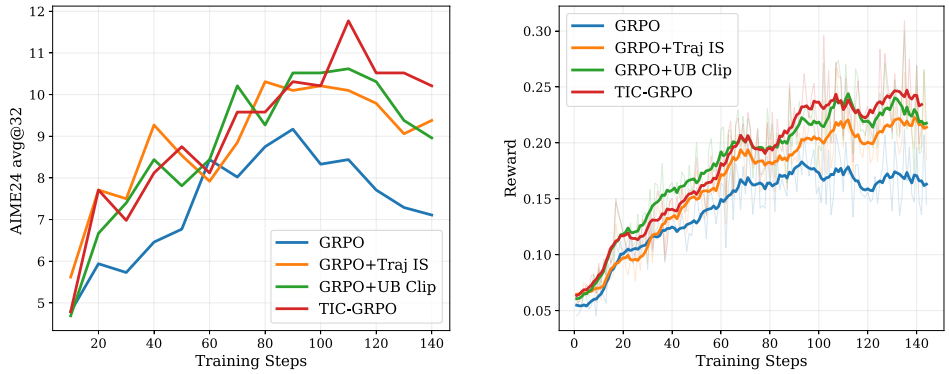
A.3 Ablation study

We evaluate component variants of TIC-GRPO on Qwen3-1.7B, using the experimental GRPO baseline for comparison. The variants use the mechanisms introduced in Section 3:

- **GRPO+UB_Clip**: uses the proposed **up-only clipping mechanism** with sampled-token normalization. It retains separate token-level importance weights and caps their upper tails for either sign of the group-relative advantage. This variant evaluates upper-tail control while retaining token-level weighting.
- **GRPO+Traj_IS**: replaces the token-level importance sampling used in GRPO with **trajectory-level importance sampling**, while keeping the original optimization procedure unchanged. The full-response likelihood ratio supplies a shared importance weight to the active-token scores within a response, allowing us to assess trajectory-level correction separately from up-only clipping.
- **TIC-GRPO**: combines up-only clipping with trajectory-level importance sampling. The upper cap acts on the shared trajectory weight, so this variant brings both components into the same update. Its comparison with each individual variant measures the empirical benefit of their combination.

Table 2: Ablation study of TIC-GRPO on Qwen3-1.7B. We compare the GRPO baseline with two component variants—up-only clipping (GRPO+UB_Clip) and trajectory-level importance sampling (GRPO+Traj_IS)—as well as their combination (TIC-GRPO).

Method	**AIME24** (Avg@32)	**AIME25** (Avg@32)	**MATH500** (Avg@1)	**Live-CodeBench** (Avg@1)
		Qwen3-1.7B		
GRPO	9.17	5.31	66.6	10.67
GRPO+UB_Clip	10.31 (+1.14)	6.64 (+1.33)	67.4 (+0.8)	11.34 (+0.67)
GRPO+Traj_IS	10.62 (+1.45)	6.77 (+1.46)	68.0 (+1.4)	11.44 (+0.77)
TIC-GRPO	**11.77 (+2.60)**	**6.98 (+1.67)**	**69.8 (+3.2)**	**12.11 (+1.44)**



(a) AIME24 Avg@32 accuracy curves during training.

(b) Corresponding training reward curves.

Figure 3: Ablation analysis of TIC-GRPO on Qwen3-1.7B.

Table 2 reports the benchmark scores, and Figure 3 shows the corresponding AIME24 accuracy and training reward curves. Both component variants improve on the GRPO baseline across all four benchmarks. The trajectory-level variant exceeds the up-only clipping variant, attaining scores of 10.62, 6.77, 68.0, and 11.44 on AIME24, AIME25, MATH500, and LiveCodeBench, respectively. Combining the two components raises these scores to 11.77, 6.98, 69.8, and 12.11. Relative to the stronger individual variant, TIC-GRPO therefore gains 1.15 points on AIME24, 0.21 on AIME25, 1.8 on MATH500, and 0.67 on LiveCodeBench. The combined method leads on both mathematical reasoning and coding evaluations, supporting the complementary practical effects of upper-tail clipping and trajectory-level importance correction.

A.4 Update magnitudes including GSPO

We compare the DAPO-style GRPO baseline, GRPO_2, GSPO, and TIC-GRPO on Qwen3-0.6B and Qwen3-1.7B using matched data, optimizer, and training settings as described in Appendix A.1. At each of the first 200 optimizer steps, we record $\|\theta_{k+1} - \theta_k\|_2^2$, the squared Euclidean distance between the parameter iterates immediately before and after that step. Table 3 reports its mean within each of eight consecutive 25-step intervals, followed by its mean and temporal standard deviation across all 200 steps. The standard deviation summarizes variation across optimizer steps within a run.

The interval summaries follow the optimization process: the first row captures initial updates, while successive rows track how update magnitudes evolve as the policy changes. Because all intervals contain the same number of steps, they contribute equally to the overall mean. The overall temporal standard deviation is computed from individual steps and therefore also reflects fluctuations within each interval. We compare methods within each model scale under the matched setup. Together, the mean and standard deviation describe the typical squared parameter displacement and the stability of its magnitude over the training window.

Table 3: Squared Euclidean distances between consecutive parameter iterates, in units of 10^6. Interval rows report temporal means; each overall row reports the mean $\pm$ temporal standard deviation across optimizer steps 1–200.

Optimizer steps	GRPO	GRPO_2	GSPO	TIC-GRPO
		Qwen3-0.6B		
1–25	26.33	31.63	26.93	23.54
26–50	45.29	52.39	45.02	29.09
51–75	48.37	40.25	41.40	28.02
76–100	32.09	31.88	31.81	26.54
101–125	30.70	30.42	28.91	25.50
126–150	32.51	25.96	33.17	25.63
151–175	25.46	25.76	28.77	26.06
176–200	29.73	30.17	30.39	25.97
Overall	33.81 ± 11.14	33.56 ± 10.66	33.30 ± 9.03	26.29 ± 2.78
		Qwen3-1.7B		
1–25	88.29	85.75	98.23	78.71
26–50	143.71	172.75	150.52	91.61
51–75	139.63	163.99	168.42	84.64
76–100	128.41	118.50	117.31	81.83
101–125	117.28	88.70	88.94	77.17
126–150	102.50	87.26	81.86	78.69
151–175	94.98	95.76	76.61	76.94
176–200	94.46	86.62	92.60	75.63
Overall	113.66 ± 26.57	112.41 ± 37.90	109.31 ± 37.18	80.65 ± 8.86

TIC-GRPO has the smallest overall mean and temporal standard deviation among the four methods at both model scales. Relative to GSPO, its mean is lower by about 21% on Qwen3-0.6B and 26% on Qwen3-1.7B, while its temporal standard deviation is lower by about 69% and 76%, respectively. The interval means also show the smaller update magnitudes during steps 26–75: GSPO reaches 45.02 and 41.40 on Qwen3-0.6B, versus 29.09 and 28.02 for TIC-GRPO; on Qwen3-1.7B, the corresponding values are 150.52 and 168.42 for GSPO, versus 91.61 and 84.64 for TIC-GRPO.

B Score moments and objective geometry

Lemma B.1 (Score moments)**.** *Under Assumption 4.1, at every active prefix,*

$$\|g_t^\theta\|^2 \le -2L \log \mathbb{P}_\theta(s_t \mid s_{t-1}), \tag{15}$$

$$\left\|\mathbb{E}_\theta[g_t^\theta (g_t^\theta)^\top \mid s_{t-1}]\right\|_{\text{op}} \le L, \qquad \mathbb{E}_\theta[\|g_t^\theta\|^2 \mid s_{t-1}] \le 2L \log |\mathcal{V}|. \tag{16}$$

Consequently, for each prompt,

$$\left\|\mathbb{E}_\theta[G_T^\theta (G_T^\theta)^\top \mid c]\right\|_{\text{op}} \le L\, \mathbb{E}_\theta[|s_T| \mid c], \qquad \mathbb{E}_\theta[\|G_T^\theta\|^2 \mid c] \le 2L \log |\mathcal{V}|\, \mathbb{E}_\theta[|s_T| \mid c]. \tag{17}$$

Proof. The negative log probability is nonnegative and has an L-Lipschitz gradient on the parameter space. The descent inequality evaluated one gradient step of length $1/L$ away gives (15) when $L > 0$. If $L = 0$, a globally affine nonnegative negative log probability has zero gradient, so every asserted inequality is immediate.

For the directional Fisher estimate, restrict the finite next-token probability vector to any line in parameter space. Its log derivatives are Lipschitz. Differentiating the normalization identity twice at the almost-everywhere points of second differentiability shows that the expected squared directional score is the negative expected directional second derivative of the log probability. The latter is bounded by L times the squared norm of the direction. Continuity of the finite sum of squared scores extends this bound to every point of the line. This proves the operator-norm assertion. Taking the

expectation in (15) and using the entropy bound for a distribution on $|\mathcal{V}|$ symbols proves the trace assertion.

The conditional mean of g_t^θ is zero, since the derivatives of the next-token probabilities sum to zero. For $i < j$, g_i^θ is measurable before token j, so conditioning on that prefix gives

$$\mathbb{E}_\theta[g_i^\theta (g_j^\theta)^\top \mid c] = 0.$$

The active-token indicator is determined by the preceding prefix. Thus the diagonal terms in the second moment of the stopped sum are bounded by the one-token estimates times the expected number of active transitions. This proves both inequalities in Eq. (17) and Property 4.1. The action that terminates the response is included; only subsequent padding transitions have zero score. □

Lemma B.2 (Stopped normalized score)**.** *For each prompt,*

$$\left\| \mathbb{E}_\theta \left[\frac{G_T^\theta (G_T^\theta)^\top}{|s_T|^2} \,\middle|\, c \right] \right\|_{\mathrm{op}} \le L \sum_{t=1}^{T} \frac{1}{t^2}. \tag{18}$$

Its trace is at most $2L \log |\mathcal{V}| \sum_{t=1}^{T} t^{-2}$.

Proof. Telescope the squared projection of the partial score sum divided by the square of its time index, stopping at $|s_T|$. The mixed term with the next increment has zero conditional mean. The coefficient of the preceding squared projection is $t^{-2} - (t-1)^{-2} \le 0$ for $t \ge 2$. After treating $t = 1$ separately, the directional conditional bound (16) gives (18). The event that token t is active is measurable before that token is sampled, so the stopping indicators preserve the conditional calculation. The same telescoping argument applied to the squared norm and the trace bound proves the trace assertion. □

Lemma B.3 (Reward geometry)**.** *For all parameters* θ, θ'*,*

$$\|\nabla J(\theta)\|^2 \le R^2 L\, \mathbb{E}_\theta[|s_T|] \le R^2 LT, \qquad \|\nabla J(\theta') - \nabla J(\theta)\| \le 2RLT \|\theta' - \theta\|. \tag{19}$$

Proof. The policy-gradient identity and the directional form of the first inequality in Eq. (17) give the first bound in Eq. (19) by Cauchy–Schwarz, using $|r| \le R$ and then averaging over prompts.

The trajectory score is $L|s_T|$-Lipschitz. Along any line segment, differentiating the reward gradient almost everywhere gives the expected reward times the trajectory-score derivative plus the expected reward times the outer product of the trajectory score applied to the parameter direction. Their expected absolute projections onto a unit direction are bounded, respectively, by $RL\,\mathbb{E}_\theta\,|s_T|$ and $RL\,\mathbb{E}_\theta\,|s_T|$, times the norm of the parameter direction. The first uses score Lipschitzness; the second uses directional Fisher and Cauchy–Schwarz. Integration proves the second bound in Eq. (19). Finite response supports justify the calculation for each prompt, and the uniform bound $2RLT$ permits averaging over prompts. Derivatives along the segment suffice; ambient second differentiability at every point is not required. □

C Empirical covariance and block convergence

Theorem C.1 (TIC-GRPO and GRPO_2 convergence)**.** *Under Assumptions 4.1 and 4.2 and the conditional sampling protocol, let* $N \ge 2$*,* $K \ge 1$*,* $B \ge 1$*,* $|G| \ge 2$*, and* $|\mathcal{V}| \ge 2$*. For* $0 < \eta \le (128RLK)^{-1}$*, TIC-GRPO satisfies*

$$\begin{aligned}
&\frac{1}{(N-1)K} \sum_{n=1}^{N-1} \sum_{k=0}^{K-1} \mathbb{E}\, \|\nabla J(\theta_{n,k})\|^2 \\
&\le \frac{24T[J^* - J(\theta_{1,0})]}{\eta K(N-1)} \\
&\quad + \left(1 - \frac{1}{|G|}\right)^2 R^2 L c_+ T \left(\frac{1}{B} + \frac{32 \log |\mathcal{V}|}{|G|}\right) \left[48RLK\eta + 392R^2L^2K(K-1)\eta^2\right].
\end{aligned} \tag{20}$$

Under the same conditions, GRPO$_2$ satisfies

$$\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla J(\theta_{n,k})\right\|^2 \leq \frac{24T[J^*-J(\theta_{1,0})]}{\eta K(N-1)} + \left(1-\frac{1}{|G|}\right)^2 R^2 L c_+ T^2\left(\frac{1}{B}+\frac{32\log|\mathcal{V}|}{|G|}\right)\left[48RLK\eta+392R^2L^2K(K-1)\eta^2\right]. \quad (21)$$

Each method selects its own constant step size. For a positive initial gap, minimizing its respective displayed bound over $0<\eta\leq(128RLK)^{-1}$ yields, for TIC-GRPO,

$$\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla J(\theta_{n,k})\right\|^2 \leq \frac{6144R^2LT}{N-1}+102\left(1-\frac{1}{|G|}\right)R^2LT\sqrt{\frac{c_+}{N-1}\left(\frac{1}{B}+\frac{32\log|\mathcal{V}|}{|G|}\right)}, \quad (22)$$

and, for GRPO$_2$,

$$\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\nabla J(\theta_{n,k})\right\|^2 \leq \frac{6144R^2LT}{N-1}+102\left(1-\frac{1}{|G|}\right)R^2LT\sqrt{\frac{c_+T}{N-1}\left(\frac{1}{B}+\frac{32\log|\mathcal{V}|}{|G|}\right)}. \quad (23)$$

The same two bounds hold for a zero initial gap with the explicit positive step choices given below.

All inner-step conditional expectations in this appendix are given $\mathscr{F}_{n,k}$. In particular, the current and old parameters are fixed before drawing the next batch. Token sums include only active transitions, and the normalizing denominator is the deterministic maximum length T.

C.1 Moments of the capped score features

For the TIC feature, the inequality $(\rho^+_{0:T})^2\leq c_+\rho_{0:T}$ changes the entire response law. Lemma B.1 then gives, for every prompt,

$$\left\|\mathbb{E}_{\theta_{n,0}}\left[(\rho^+_{0:T})^2 G_T^{\theta_{n,k}}(G_T^{\theta_{n,k}})^\top \mid c\right]\right\|_{\mathrm{op}}\leq c_+LT,$$
$$\mathbb{E}_{\theta_{n,0}}\left[(\rho^+_{0:T})^2\|G_T^{\theta_{n,k}}\|^2 \mid c\right]\leq 2c_+LT\log|\mathcal{V}|. \quad (24)$$

The current-policy trajectory-score identity is applied after the change of measure. It applies to the score before multiplication by the reward or the empirical advantage.

For GRPO$_2$, Cauchy–Schwarz for the T token positions gives the positive-semidefinite matrix inequality

$$\left(\sum_{t=1}^T \rho_t^+ g_t^{\theta_{n,k}}\right)\left(\sum_{t=1}^T \rho_t^+ g_t^{\theta_{n,k}}\right)^\top \preceq T\sum_{t=1}^T(\rho_t^+)^2 g_t^{\theta_{n,k}}(g_t^{\theta_{n,k}})^\top.$$

At an active old-policy prefix, $(\rho_t^+)^2\leq c_+\rho_t$ changes only the next-token law. The directional bound in (16) gives an operator norm at most c_+L for each term; its trace is at most $2c_+L\log|\mathcal{V}|$. Summing the T positions proves

$$\left\|\mathbb{E}_{\theta_{n,0}}\left[\left(\sum_{t=1}^T \rho_t^+ g_t^{\theta_{n,k}}\right)\left(\sum_{t=1}^T \rho_t^+ g_t^{\theta_{n,k}}\right)^\top \mid c\right]\right\|_{\mathrm{op}}\leq c_+LT^2,$$
$$\mathbb{E}_{\theta_{n,0}}\left[\left\|\sum_{t=1}^T \rho_t^+ g_t^{\theta_{n,k}}\right\|^2 \mid c\right]\leq 2c_+LT^2\log|\mathcal{V}|. \quad (25)$$

This calculation bounds the complete token sum, including its cross terms, by a sum of single-token second moments.

C.2 The empirical baseline and the batch variance

The empirical covariance in a response group has expectation $1 - |G|^{-1}$ times the population reward-feature covariance. To verify this, expand the empirical reward mean: among its $|G|^2$ products, $|G|$ contain a reward and feature from the same response, while the other $|G|(|G| - 1)$ contain independent responses from that prompt. Consequently,

$$\mathbb{E}[\widehat{G}^{(n,k)}_{\mathrm{TIC}} \mid \mathscr{F}_{n,k}] = \frac{1 - |G|^{-1}}{T} \mathbb{E}_{\theta_{n,0}}\left[(r(s_T) - \mu_{\theta_{n,0},c})\rho^+_{0:T} G^{\theta_{n,k}}_T\right],$$
$$\mathbb{E}[\widehat{G}^{(n,k)}_{\mathrm{GRPO}_2} \mid \mathscr{F}_{n,k}] = \frac{1 - |G|^{-1}}{T} \mathbb{E}_{\theta_{n,0}}\left[(r(s_T) - \mu_{\theta_{n,0},c}) \sum_{t=1}^{T} \rho^+_t g^{\theta_{n,k}}_t\right]. \tag{26}$$

The expectation on the right includes the prompt draw.

Here is a finite-sample variance calculation retaining the same-group baseline. Write the empirical covariance as $1 - |G|^{-1}$ times the average over unordered pairs of distinct responses of the product of the reward difference and the feature difference, divided by two. Each pair term has squared-norm expectation at most $4R^2$ times the single-response feature second moment: the reward difference is at most $2R$, and the expected squared feature difference is at most four times that moment. Disjoint pairs are independent. For a fixed pair, precisely $2|G| - 3$ pairs, including itself, intersect it. Cauchy–Schwarz for the covariance of intersecting pairs and

$$\frac{2|G| - 3}{\binom{|G|}{2}} \le \frac{4}{|G|}$$

therefore bound the group variance by $16(1 - |G|^{-1})^2 R^2/|G|$ times the single-response feature second moment. This also covers $|G| = 2$.

The squared norm of each population reward-feature covariance is at most R^2 times the operator norm of the feature second-moment matrix. Indeed, project onto an arbitrary unit direction, apply Cauchy–Schwarz to the centered reward and the projected feature, and use $\mathrm{Var}(r \mid c) \le R^2$. The bound is uniform over directions, so it bounds the norm of the mean vector without a trace factor.

Condition on the complete list of sampled prompt occurrences. The conditional mean of the batch is the average of the corresponding population group means. Independence of the B prompt draws bounds the variance of that average by B^{-1} times the preceding squared-mean bound. For the remaining within-group fluctuations, Jensen's inequality across the B groups bounds the conditional squared norm of their average by the average of their conditional squared norms. The latter step permits arbitrary dependence between groups. The two fluctuations are orthogonal in expectation because the second has conditional mean zero given the prompt list. Applying the group calculation and (24)–(25), and dividing by T^2, yields

$$\mathbb{E}\left[\left\|\widehat{G}^{(n,k)}_{\mathrm{TIC}} - \mathbb{E}[\widehat{G}^{(n,k)}_{\mathrm{TIC}} \mid \mathscr{F}_{n,k}]\right\|^2 \middle| \mathscr{F}_{n,k}\right] \le \frac{(1 - |G|^{-1})^2 R^2 c_+ L}{T}\left(\frac{1}{B} + \frac{32 \log|\mathcal{V}|}{|G|}\right), \tag{27}$$
$$\mathbb{E}\left[\left\|\widehat{G}^{(n,k)}_{\mathrm{GRPO}_2} - \mathbb{E}[\widehat{G}^{(n,k)}_{\mathrm{GRPO}_2} \mid \mathscr{F}_{n,k}]\right\|^2 \middle| \mathscr{F}_{n,k}\right] \le (1 - |G|^{-1})^2 R^2 c_+ L\left(\frac{1}{B} + \frac{32 \log|\mathcal{V}|}{|G|}\right). \tag{28}$$

C.3 Stability of the actual mean update

At $\theta_{n,k} = \theta_{n,0}$, both ratios equal one, so both expectations in (26) equal $(1 - |G|^{-1})\nabla J(\theta_{n,0})/T$. We bound the change of these complete population means along the parameter segment from $\theta_{n,0}$ to $\theta_{n,k}$, with the old response law and its reward mean held fixed.

For TIC, the derivative of the capped score feature has two terms: the capped ratio times the directional derivative of the trajectory score, and, where the ratio is below the cap, the ratio times the trajectory-score outer product applied to the parameter direction. For each fixed unit projection direction, the old-law expected absolute projection of the first term is bounded by LT times the norm of the parameter direction. This uses the trajectory-score Lipschitz constant $L|s_T|$, the inequality $\rho^+_{0:T} \le \rho_{0:T}$, and a full change of measure. The second term has the same bound, by the first inequality in Eq. (17) and Cauchy–Schwarz under the current response law.

For GRPO_2, apply this derivative calculation separately at every active old-policy prefix. The first term, the capped ratio times the token-score derivative, has expected absolute projection at most L times the norm of the parameter direction. In the second term, the ratio derivative is zero above the cap; below the cap, a next-token change of measure and (16) give the same bound. Summing at most T active tokens bounds the expected absolute projection of the complete feature derivative by $2LT$ times that norm. Multiplication by $|r - \mu_{\theta_{n,0},c}| \le 2R$ is valid even though the reward depends on the entire response.

These derivative calculations hold almost everywhere along the segment. Lipschitz scores are absolutely continuous there; each probability and its capped ratio are locally absolutely continuous. On a level set where the ratio equals the cap, its segment derivative vanishes almost everywhere. Finite next-token and response supports justify each conditional differentiation. The bounds just obtained are uniform over prompts and along the segment, so integration and prompt averaging are legitimate. For either $\mathsf{A} \in \{\mathrm{TIC}, \mathrm{GRPO}_2\}$ they prove

$$\left\| \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}] - \frac{1 - |G|^{-1}}{T} \nabla J(\theta_{n,0}) \right\| \le 4RL \|\theta_{n,k} - \theta_{n,0}\|. \tag{29}$$

This estimate holds also at $c_+ = 1$.

Decomposing a second moment into variance and squared mean, applying (29), and using $(1 - |G|^{-1})^2 \le 1$ gives

$$\begin{aligned} T\,\mathbb{E}[\|\widehat{G}_{\mathsf{A}}^{(n,k)}\|^2 \mid \mathscr{F}_{n,k}] \le \frac{4}{T} \|\nabla J(\theta_{n,0})\|^2 + 32R^2L^2T \|\theta_{n,k} - \theta_{n,0}\|^2 \\ + T\,\mathbb{E}\left[\left\| \widehat{G}_{\mathsf{A}}^{(n,k)} - \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}] \right\|^2 \middle| \mathscr{F}_{n,k} \right]. \end{aligned} \tag{30}$$

C.4 Block ascent and all inner iterates

For either method, the update recursion implies the pathwise bound

$$T\|\theta_{n,k} - \theta_{n,0}\|^2 \le \eta^2 k \sum_{j=0}^{k-1} T\|\widehat{G}_{\mathsf{A}}^{(n,j)}\|^2.$$

Summing over k and noting that each earlier moment has coefficient at most $K(K-1)/2$ gives

$$\sum_{k=0}^{K-1} \mathbb{E}[T\|\theta_{n,k} - \theta_{n,0}\|^2] \le \eta^2 \frac{K(K-1)}{2} \sum_{k=0}^{K-1} \mathbb{E}[T\|\widehat{G}_{\mathsf{A}}^{(n,k)}\|^2]. \tag{31}$$

Since $RLK\eta \le 1/128$, the coefficient $32R^2L^2\eta^2K(K-1)/2$ is at most $1/1024$. Substitute (31) into the sum of (30) and absorb the displacement contribution on the left. The coefficient remaining on the left is at least $1 - 1/1024$, whose reciprocal is at most two. The result is

$$\begin{aligned} \sum_{k=0}^{K-1} \mathbb{E}[T\|\widehat{G}_{\mathsf{A}}^{(n,k)}\|^2] \le \frac{8K}{T}\,\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2 \\ + 2T \sum_{k=0}^{K-1} \mathbb{E}\left\| \widehat{G}_{\mathsf{A}}^{(n,k)} - \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}] \right\|^2. \end{aligned} \tag{32}$$

These quantities are finite by the uniform feature second moments and the finite update recursion.

For $|G| \ge 2$, Young's inequality and (29) give

$$\begin{aligned} \left\langle \nabla J(\theta_{n,0}), \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}] \right\rangle \\ \ge \frac{\|\nabla J(\theta_{n,0})\|^2}{4T} - 16R^2L^2T\|\theta_{n,k} - \theta_{n,0}\|^2. \end{aligned} \tag{33}$$

In detail, the leading coefficient is at least $1/(2T)$, and $4RL\|\nabla J(\theta_{n,0})\|\|\theta_{n,k} - \theta_{n,0}\|$ is at most $\|\nabla J(\theta_{n,0})\|^2/(4T) + 16R^2L^2T\|\theta_{n,k} - \theta_{n,0}\|^2$.

Apply the $2RLT$ smoothness of J to the complete block update. Cauchy–Schwarz across its K steps, followed by (33) and (31), yields

$$\begin{aligned}&\mathbb{E}[J(\theta_{n+1,0}) - J(\theta_{n,0})]\\ &\geq \frac{\eta K}{4T}\,\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2 - \eta\left[RLK\eta + 8R^2L^2K(K-1)\eta^2\right]\sum_{k=0}^{K-1}\mathbb{E}[T\|\widehat{G}_{\mathsf{A}}^{(n,k)}\|^2].\end{aligned} \tag{34}$$

The bracket is at most $1/64$. Inserting (32) leaves a gradient coefficient at least $\eta K/(8T)$. Sum over outer blocks and telescope J, using $J(\theta_{N,0}) \leq J^*$. This proves

$$\begin{aligned}\frac{1}{N-1}\sum_{n=1}^{N-1}\frac{\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2}{T} &\leq \frac{8[J^* - J(\theta_{1,0})]}{\eta K(N-1)}\\ &+ \frac{16T[RLK\eta + 8R^2L^2K(K-1)\eta^2]}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\widehat{G}_{\mathsf{A}}^{(n,k)} - \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}]\right\|^2.\end{aligned} \tag{35}$$

Finally, objective-gradient Lipschitzness implies

$$\|\nabla J(\theta_{n,k})\|^2 \leq 2\|\nabla J(\theta_{n,0})\|^2 + 8R^2L^2T^2\|\theta_{n,k} - \theta_{n,0}\|^2.$$

Using (31) and (32) to sum the second term gives

$$\begin{aligned}&\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\,\|\nabla J(\theta_{n,k})\|^2\\ &\leq \frac{2 + 32R^2L^2K(K-1)\eta^2}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2\\ &\quad + \frac{8R^2L^2T^2K(K-1)\eta^2}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\widehat{G}_{\mathsf{A}}^{(n,k)} - \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}]\right\|^2.\end{aligned}$$

The anchor-gradient coefficient $2 + 32R^2L^2K(K-1)\eta^2$ is at most three. Substitution of (35) therefore gives

$$\begin{aligned}&\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\,\|\nabla J(\theta_{n,k})\|^2\\ &\leq \frac{24T[J^* - J(\theta_{1,0})]}{\eta K(N-1)}\\ &\quad + \frac{T^2[48RLK\eta + 392R^2L^2K(K-1)\eta^2]}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\left\|\widehat{G}_{\mathsf{A}}^{(n,k)} - \mathbb{E}[\widehat{G}_{\mathsf{A}}^{(n,k)} \mid \mathscr{F}_{n,k}]\right\|^2.\end{aligned} \tag{36}$$

The coefficient 392 equals $3 \cdot 16 \cdot 8 + 8$. Inserting (27) proves (20), and inserting (28) proves (21).

C.5 Separate step-size optimization

For a positive initial gap, each right-hand side in (20) and (21) is strictly convex in $\eta > 0$, being a positive multiple of $1/\eta$ plus a positive linear term and a nonnegative quadratic term. It diverges at both ends of this interval. For TIC, its unique positive stationary point solves

$$\begin{aligned}\frac{24[J^* - J(\theta_{1,0})]}{K(N-1)} = \left(1 - \frac{1}{|G|}\right)^2 R^2Lc_+\left(\frac{1}{B} + \frac{32\log|\mathcal{V}|}{|G|}\right)\\ \times\left[48RLK\eta^2 + 784R^2L^2K(K-1)\eta^3\right].\end{aligned}$$

The GRPO_2 equation has one additional factor T on the right. For each algorithm, the minimizer over the sufficient step interval is its own stationary point capped at $(128RLK)^{-1}$.

To obtain closed bounds with explicit parameter dependence, observe on this interval that

$$48RLK\eta + 392R^2L^2K(K-1)\eta^2 \leq 52RLK\eta.$$

For TIC choose the admissible witness

$$\eta = \frac{1}{RLK} \min\left\{ \frac{1}{128}, \sqrt{\frac{24[J^* - J(\theta_{1,0})]}{52R(N-1)(1-|G|^{-1})^2 c_+ (B^{-1} + 32\log|\mathcal{V}|/|G|)}} \right\}.$$

For GRPO$_2$, insert an additional factor T in the denominator inside the square root. If the square-root choice is admissible, the inverse-step and linear terms are equal. If the cap is active, the linear term at the cap is at most the geometric mean of their two coefficients. These two cases show that the respective minimized bounds are at most

$$\frac{3072RLT[J^* - J(\theta_{1,0})]}{N-1} + 72\left(1 - \frac{1}{|G|}\right) R^{3/2} LT \sqrt{\frac{c_+[J^* - J(\theta_{1,0})]}{N-1}\left(\frac{1}{B} + \frac{32\log|\mathcal{V}|}{|G|}\right)}$$

for TIC, and the same expression with one additional T inside the square root for GRPO$_2$. Here $2\sqrt{24 \cdot 52} < 72$. Using $J^* - J(\theta_{1,0}) \le 2R$ and $72\sqrt{2} < 102$ proves (22)–(23). For a zero initial gap, replace the gap by $2R$ in the witness step choices. They are then strictly positive and give the same uniform bounds directly. All step choices refer to the respective algorithm's own run and deterministic bound.

D Per-response normalization and GRPO

Theorem D.1 (GRPO convergence)**.** *Under Assumptions 4.1–4.2 and the conditional sampling protocol, let $N \ge 2$, $B \ge 1$, $|G| \ge 2$, $\eta > 0$, and $\epsilon_{low}, \epsilon_{high} > 0$. For $K \ge 2$, original GRPO satisfies*

$$\begin{aligned}
&\frac{1}{(N-1)K} \sum_{n=1}^{N-1} \sum_{k=0}^{K-1} \mathbb{E}\,\|\nabla J(\theta_{n,k})\|^2 \\
&\le \frac{4T\big(J^* - J(\theta_{1,0})\big)}{\eta(N-1)K} + \frac{16R^3L^2T^2\eta}{\mathrm{e}(N-1)} \sum_{n=1}^{N-1} \mathbb{E}[\mathcal{M}_n^2] \\
&\quad + 8TR^2(LT)^{3/2}\eta\left(3 + \max\left\{1, \frac{3}{2\min\{\epsilon_{low}, \epsilon_{high}\}}\right\}\right)\sqrt{\frac{(K-1)(2K-1)}{6}} \\
&\qquad\qquad \times \sqrt{\frac{4R^2L}{\mathrm{e}(N-1)} \sum_{n=1}^{N-1} \mathbb{E}[\mathcal{M}_n^2]} \\
&\quad + \frac{8R^2LT}{N-1}\left(\sum_{t=1}^{T} \frac{1}{t^2}\right) \sum_{n=1}^{N-1} \mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right].
\end{aligned} \tag{37}$$

The inverse-probability moments are evaluated along the run at step size η. The bound is interpreted in the extended real sense, allowing infinite inverse-probability moments. For $K = 1$, every update is on policy and

$$\begin{aligned}
&\frac{1}{N-1} \sum_{n=1}^{N-1} \mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2 \\
&\le \frac{4T[J^* - J(\theta_{1,0})]}{\eta(N-1)} + 8R^3L^2T^2\eta\log|\mathcal{V}| \sum_{t=1}^{T} t^{-2} \\
&\quad + \frac{8R^2LT}{N-1}\left(\sum_{t=1}^{T} t^{-2}\right) \sum_{n=1}^{N-1} \mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right].
\end{aligned} \tag{38}$$

For a positive initial gap, the step size

$$\eta = \sqrt{\frac{J^* - J(\theta_{1,0})}{2R^3L^2T(N-1)\log|\mathcal{V}| \sum_{t=1}^{T} t^{-2}}}$$

minimizes the first two terms of (38)*, giving*

$$\frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2$$
$$\le 8\sqrt{2}\,R^{3/2}LT^{3/2}\sqrt{\frac{[J^*-J(\theta_{1,0})]\log|\mathcal{V}|\sum_{t=1}^{T}t^{-2}}{N-1}}$$
$$+\frac{8R^2LT}{N-1}\left(\sum_{t=1}^{T}t^{-2}\right)\sum_{n=1}^{N-1}\mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right]. \tag{39}$$

For zero initial gap, $\eta = 1/\sqrt{T(N-1)}$ *gives*

$$\frac{1}{N-1}\sum_{n=1}^{N-1}\mathbb{E}\,\|\nabla J(\theta_{n,0})\|^2$$
$$\le \frac{8R^3L^2T^{3/2}\log|\mathcal{V}|}{\sqrt{N-1}}\sum_{t=1}^{T}t^{-2}$$
$$+\frac{8R^2LT}{N-1}\left(\sum_{t=1}^{T}t^{-2}\right)\sum_{n=1}^{N-1}\mathbb{E}\left[\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right]. \tag{40}$$

Each length term is evaluated along the run generated by its respective step choice. If $|\mathcal{V}|=1$*, the objective gradient and initial gap are zero, so only the zero-gap branch applies.*

Away from a clipping threshold, the multiplier in the first expression of Eq. (9) is

$$m_t^{\mathrm{GRPO}} = \begin{cases}\rho_t\,\mathbf{1}\{\rho_t \le 1+\epsilon_{\text{high}}\}, & A_c(s_T)\ge 0,\\ \rho_t\,\mathbf{1}\{\rho_t \ge 1-\epsilon_{\text{low}}\}, & A_c(s_T)<0.\end{cases}$$

At a threshold the multiplier is any value between the adjacent one-sided values. Thus $0 \le m_t^{\mathrm{GRPO}} \le \rho_t$, including thresholds. The algorithm averages Z_{GRPO} over the actual prompt and response weights.

Lemma D.1 (The old-policy GRPO mean)**.** *For positive clipping margins,*

$$\mathbb{E}\left[\widehat{G}^{(k)}_{\mathrm{GRPO}}(\theta_{n,0},\theta_{n,0})\middle|\mathscr{F}_{n,k}\right]$$
$$=\left(1-\frac{1}{|G|}\right)\frac{\nabla J(\theta_{n,0})}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}$$
$$+\left(1-\frac{1}{|G|}\right)\mathbb{E}_{\theta_{n,0}}\left[\left(\frac{1}{|s_T|}-\frac{1}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}\right)(r(s_T)-\mu_{\theta_{n,0},c})G_T^{\theta_{n,0}}\right]. \tag{41}$$

The squared norm of the last expectation, including its prefactor, is at most

$$4\left(1-\frac{1}{|G|}\right)^2 R^2L\left(\sum_{t=1}^{T}\frac{1}{t^2}\right)\frac{\sigma^2_{\theta_{n,0}}}{(\mathbb{E}_{\theta_{n,0}}[|s_T|])^2}. \tag{42}$$

Proof. At the old parameter every ratio equals one and lies in the interior of each active clipping branch. Expand the empirical reward mean in each group's update. Among the $|G|^2$ pairs of a reward and a trajectory score divided by length, $|G|$ pairs use the same response and $|G|(|G|-1)$ use independent responses from the same prompt. This gives exactly

$$\left(1-\frac{1}{|G|}\right)\mathbb{E}_{\theta_{n,0}}\left[(r(s_T)-\mu_{\theta_{n,0},c})\frac{G_T^{\theta_{n,0}}}{|s_T|}\right].$$

Adding and subtracting the reciprocal joint mean length gives (41), because the old trajectory score has conditional mean zero given its prompt. For the remainder, write the reciprocal difference as

$$\frac{\mathbb{E}_{\theta_{n,0}}[|s_T|]-|s_T|}{|s_T|\,\mathbb{E}_{\theta_{n,0}}[|s_T|]}.$$

In any direction, Cauchy–Schwarz, $|r-\mu_{\theta_{n,0},c}|\le 2R$, and (18) give (42). Both the expectation and the variance are taken jointly over prompts and responses. □

Lemma D.2 (GRPO drift and second moment)**.** *For the same empirical group and positive clipping margins,*

$$\left\|\mathbb{E}\left[\widehat{G}^{(k)}_{\mathrm{GRPO}}(\theta_{n,k},\theta_{n,0})-\widehat{G}^{(k)}_{\mathrm{GRPO}}(\theta_{n,0},\theta_{n,0})\middle|\mathscr{F}_{n,k}\right]\right\|$$
$$\leq 2RLT\left(2+\max\left\{1,\frac{3}{2\min\{\epsilon_{low},\epsilon_{high}\}}\right\}\right)\|\theta_{n,k}-\theta_{n,0}\|, \tag{43}$$
$$\left\|\widehat{G}^{(n,k)}_{\mathrm{GRPO}}\right\|^2\leq\frac{4R^2L}{\mathrm{e}}\mathcal{M}_n^2. \tag{44}$$

Proof. At a fixed active prefix, integrate the ratio-score product along the segment from $\theta_{n,0}$ to $\theta_{n,k}$. Its derivative is the sum of the ratio times the score derivative and the ratio times the outer product of the score applied to the parameter displacement. In the conditional old-policy expectation of the absolute projection onto any direction, change the next-action law using the ratio. The Lipschitz score bound controls the first term by $L\|\theta_{n,k}-\theta_{n,0}\|$, and the directional Fisher bound controls the second by the same quantity. Hence

$$\sup_{\|\theta'-\theta\|=1}\mathbb{E}_{\theta_{n,0}}\left[\left|\left\langle\theta'-\theta,\rho_t g_t^{\theta_{n,k}}-g_t^{\theta_{n,0}}\right\rangle\right|\middle|s_{t-1}\right]\leq 2L\|\theta_{n,k}-\theta_{n,0}\|.$$

On an inactive branch, including a threshold with a fractional active coefficient, the ratio differs from one by at least $\min\{\epsilon_{\text{low}},\epsilon_{\text{high}}\}$. Express the ratio difference as the integral of its derivative along the same segment. Markov's inequality applied pointwise to this absolute integral, followed by a next-action change of measure, Cauchy–Schwarz, and score Lipschitzness, bounds the expected absolute old-score projection on that event by

$$\frac{L\|\theta_{n,k}-\theta_{n,0}\|}{\min\{\epsilon_{\text{low}},\epsilon_{\text{high}}\}}\left(1+\frac{\sqrt{L}\|\theta_{n,k}-\theta_{n,0}\|}{2}\right).$$

When $\sqrt{L}\|\theta_{n,k}-\theta_{n,0}\|\leq 1$, this is at most the factor $3/(2\min\{\epsilon_{\text{low}},\epsilon_{\text{high}}\})$ times $L\|\theta_{n,k}-\theta_{n,0}\|$. Otherwise the unconditional old-score first-moment bound $\sqrt{L}$ is at most $L\|\theta_{n,k}-\theta_{n,0}\|$. These two bounds control the inactive part globally. The active part is bounded by the preceding ratio-score estimate. Finally use $|A_c|\leq 2R$, $|s_T|^{-1}\leq 1$, and at most T active tokens, before taking expectations. The branch-event inclusion and reward bounds hold pathwise, so the conditional estimates permit dependence of the active branch on the same group's empirical advantage. They prove (43).

For (44), self-bounding gives

$$\|\nabla\,\mathbb{P}_\theta(s_t\mid s_{t-1})\|^2\leq -2L\,\mathbb{P}_\theta(s_t\mid s_{t-1})^2\log\mathbb{P}_\theta(s_t\mid s_{t-1})\leq L/\mathrm{e}.$$

The last inequality follows by maximizing the scalar function $-x^2\log x$ on $(0,1]$, whose maximum is $1/(2\mathrm{e})$. Use $0\leq m_t^{\mathrm{GRPO}}\leq\rho_t$ and Jensen's inequality for the per-response token average and the empirical response and prompt averages. This yields the stated pathwise bound with the maximum over the sampled transitions. The resulting moment bound retains this random maximum inside the expectation. □

D.1 Proof of Theorem D.1

Proof. First let $K\geq 2$. If the displayed inverse-probability moment is infinite, (37) holds in the extended real sense. Otherwise, the moment is finite and the following estimates apply. Pair the old-policy identity (41) with the current gradient. Young's inequality assigns half of the coefficient $(1-|G|^{-1})/\,\mathbb{E}_{\theta_{n,0}}|s_T|$ to the length remainder. Use (42), (43), and the second bound in Eq. (19). Since $|G|\geq 2$ and the mean length is at most T, this gives

$$\left\langle\nabla J(\theta_{n,k}),\mathbb{E}[\widehat{G}^{(n,k)}_{\mathrm{GRPO}}\mid\mathscr{F}_{n,k}]\right\rangle\geq\frac{\|\nabla J(\theta_{n,k})\|^2}{4T}$$
$$-2R^2(LT)^{3/2}\left(3+\max\left\{1,\frac{3}{2\min\{\epsilon_{\text{low}},\epsilon_{\text{high}}\}}\right\}\right)\|\theta_{n,k}-\theta_{n,0}\|$$
$$-2R^2L\left(\sum_{t=1}^{T}\frac{1}{t^2}\right)\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]}.$$

Here the first bound in Eq. (19) controls the current-gradient norm when pairing it with the drift. The objective ascent inequality adds a cost at most $RLT\eta^2\,\mathbb{E}[\|\widehat{G}^{(n,k)}_{\mathrm{GRPO}}\|^2 \mid \mathscr{F}_{n,k}]$.

Summing the inner-update displacements and applying Cauchy–Schwarz gives

$$\begin{aligned}
&\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\,\|\theta_{n,k}-\theta_{n,0}\| \\
&\le \eta\sqrt{\frac{(K-1)(2K-1)}{6}}\sqrt{\frac{1}{(N-1)K}\sum_{n=1}^{N-1}\sum_{k=0}^{K-1}\mathbb{E}\,\|\widehat{G}^{(n,k)}_{\mathrm{GRPO}}\|^2} \\
&\le \eta\sqrt{\frac{(K-1)(2K-1)}{6}}\sqrt{\frac{4R^2L}{\mathrm{e}(N-1)}\sum_{n=1}^{N-1}\mathbb{E}\,\mathcal{M}_n^2}.
\end{aligned}$$

For the first inequality, each update with index j appears in the later displacements $K-1-j$ times; the average of the squares of these coefficients is $(K-1)(2K-1)/6$. Sum the ascent inequalities over all $(N-1)K$ updates, telescope J, bound the total increase by $J^*-J(\theta_{1,0})$, and multiply by $4T/\eta$. Substitution of (44) proves (37).

When $K=1$, every update is on policy. Empirical covariance Cauchy–Schwarz gives, pathwise,

$$\left\|\widehat{G}^{(n,0)}_{\mathrm{GRPO}}\right\|^2 \le R^2\frac{1}{B}\sum_{\text{prompt occurrences}}\sum_{s_T\in\mathcal{S}_T}\xi_c(s_T)\frac{\|G_T^{\theta_{n,0}}\|^2}{|s_T|^2}.$$

The trace form of (18) bounds its expectation by $2R^2L\log|\mathcal{V}|\sum_{t=1}^{T}t^{-2}$. Substitution into the same ascent argument proves (38).

The sum $\sum_{t=1}^{T}t^{-2}$ is at most 2, by comparison of its terms after the first with the integral of x^{-2} over $[1,\infty)$. Absorbing only R, L, K, the clipping margins, and the initial gap into constants in (37) proves (13). For $K=1$, its step-dependent part is

$$\frac{4T[J^*-J(\theta_{1,0})]}{\eta(N-1)}+8R^3L^2T^2\eta\log|\mathcal{V}|\sum_{t=1}^{T}\frac{1}{t^2}.$$

For positive initial gap, differentiation or the arithmetic–geometric mean inequality gives the step size in Theorem D.1. Its value is

$$8\sqrt{2}\,R^{3/2}LT^{3/2}\sqrt{\frac{[J^*-J(\theta_{1,0})]\log|\mathcal{V}|\sum_{t=1}^{T}t^{-2}}{N-1}}.$$

Adding the length term proves (39); that term remains evaluated along the run at the selected step size. For zero initial gap, substituting $\eta=1/\sqrt{T(N-1)}$ into (38) proves (40). Both step choices imply (14).

Finally, averaging $(|s_T|-1)(T-|s_T|)\ge 0$ under the joint old law gives

$$\begin{aligned}
\frac{\sigma^2_{\theta_{n,0}}}{\mathbb{E}_{\theta_{n,0}}[|s_T|]} &\le T+1-\mathbb{E}_{\theta_{n,0}}[|s_T|]-\frac{T}{\mathbb{E}_{\theta_{n,0}}[|s_T|]} \\
&\le (\sqrt{T}-1)^2.
\end{aligned}$$

The second inequality is the arithmetic–geometric mean inequality. Thus the individual-length contribution in (37) is at most $16R^2LT(\sqrt{T}-1)^2$, as stated in the main text. □

E Conditional core moments

Proposition E.1 (Core update moments)**.** *Under the stated sampling protocol and Assumptions 4.1–4.2,*

$$\begin{aligned}
&\mathbb{E}[(\|Z_{\mathrm{GRPO}}\|^2,\|Z_{\mathrm{GRPO}_2}\|^2,\|Z_{\mathrm{TIC}}\|^2)\mid\mathscr{F}_{n,k}] \\
&\quad\le 4R^2L\left(\frac{\mathbb{E}[\mathcal{M}_n^2\mid\mathscr{F}_{n,k}]}{\mathrm{e}},2c_+\log|\mathcal{V}|,\frac{2c_+\log|\mathcal{V}|}{T}\right).
\end{aligned}\tag{45}$$

Expectation and inequality are interpreted componentwise, in the extended real sense for the first component. Here the distinguished response is sampled using the empirical prompt and response weights. The same upper bounds apply to the empirical average update by Jensen's inequality.

Proof of Proposition E.1. For GRPO, $|A_c| \le 2R$, the inequality $0 \le m_t^{\mathrm{GRPO}} \le \rho_t$, and the probability-gradient bound in Lemma D.2 give the pathwise estimate

$$\|Z_{\mathrm{GRPO}}\|^2 \le \frac{4R^2}{|s_T|} \sum_{t=1}^{|s_T|} \frac{L}{\mathrm{e}\,\mathbb{P}_{\theta_{n,0}}(s_t \mid s_{t-1})^2} \le \frac{4R^2 L}{\mathrm{e}} \mathcal{M}_n^2.$$

Take conditional expectations to obtain the first componentwise inequality in Eq. (45).

For GRPO$_2$, the bound $|A_c| \le 2R$ and (25) give

$$\mathbb{E}[\|Z_{\mathrm{GRPO}_2}\|^2 \mid \mathscr{F}_{n,k}] \le \frac{4R^2}{T^2}\, 2c_+ LT^2 \log|\mathcal{V}| = 8R^2 c_+ L \log|\mathcal{V}|.$$

This is the second componentwise inequality in Eq. (45). For TIC, (24) similarly gives

$$\mathbb{E}[\|Z_{\mathrm{TIC}}\|^2 \mid \mathscr{F}_{n,k}] \le \frac{4R^2}{T^2}\, 2c_+ LT \log|\mathcal{V}| = \frac{8R^2 c_+ L \log|\mathcal{V}|}{T}.$$

This is the third componentwise inequality in Eq. (45). The empirical advantage is bounded pathwise before taking expectations, so it need not be independent of the score. Jensen's inequality for the nonnegative empirical weights extends the bounds to the averaged updates. □

F Hard-instance proofs for the variance bounds

Proposition F.1 (Response-level variance lower bounds)**.** *Fix $R, L > 0$ and $c_+ \ge 1$. For every $T \ge 2$ and $|G| \ge 2$, there are separate one-dimensional binary policy instances satisfying Assumptions 4.1–4.2, with one prompt, independent responses, and deterministic response length T, for which*

$$\mathbb{E}\Big[\big\|Z_{\mathrm{GRPO}_2} - \mathbb{E}[Z_{\mathrm{GRPO}_2} \mid \mathscr{F}_{n,k}]\big\|^2 \mid \mathscr{F}_{n,k}\Big] \ge \frac{R^2 L(|G|-1)}{256|G|}, \tag{46}$$

$$\mathbb{E}\Big[\big\|Z_{\mathrm{TIC}} - \mathbb{E}[Z_{\mathrm{TIC}} \mid \mathscr{F}_{n,k}]\big\|^2 \mid \mathscr{F}_{n,k}\Big] \ge \frac{R^2 L(|G|-1)}{8|G|T}. \tag{47}$$

For original GRPO, a binary instance with $|G| = 2$ satisfies, whenever $T \ge 64$ and $T^2 > 2c_+$,

$$\mathbb{E}\Big[\big\|Z_{\mathrm{GRPO}} - \mathbb{E}[Z_{\mathrm{GRPO}} \mid \mathscr{F}_{n,k}]\big\|^2 \mid \mathscr{F}_{n,k}\Big] \ge \frac{R^2 LT}{256}. \tag{48}$$

Here the cores are evaluated pointwise at fixed old/current parameter pairs before sampling. In the original-GRPO construction, the old-policy success probability is T^{-2}; in the GRPO$_2$ construction it is $1/(4c_+)$, independently of T.

Proof of Proposition F.1. Use one prompt and a scalar parameter. At every parameter-dependent transition, let the probability of token 1 be

$$\frac{\exp(2\sqrt{L}\,\theta)}{1+\exp(2\sqrt{L}\,\theta)}.$$

Its score is $2\sqrt{L}$ times the difference between the observed binary token and its success probability. The derivative of this score has absolute value

$$4L\frac{\exp(2\sqrt{L}\,\theta)}{(1+\exp(2\sqrt{L}\,\theta))^2} \le L.$$

Thus the score is globally L-Lipschitz, and both token probabilities are positive for every finite parameter. Responses have exactly T tokens. All groups consist of independent responses, and the parameter pairs below are fixed before sampling.

GRPO$_2$. Let the first $T-1$ tokens follow this policy independently. The last token is an independent, parameter-independent fair binary token, and the terminal reward is R times that token. This reward lies in $[0, R]$ and the last score is zero. Set

$$\theta_{n,k} = 0, \qquad \theta_{n,0} = -\frac{\log(4c_+ - 1)}{2\sqrt{L}}.$$

The current success probability is $1/2$, while the old success probability is $1/(4c_+)$. For $t \le T-1$, the score at the current parameter is $\sqrt{L}$ on token 1 and $-\sqrt{L}$ on token 0. The clipped ratio is c_+ on token 1 and $2c_+/(4c_+ - 1)$ on token 0. Hence, directly under the old response law,

$$\mathbb{E}[\rho_t^+ g_t^{\theta_{n,k}} \mid \mathscr{F}_{n,k}] = \frac{c_+\sqrt{L}}{4c_+} - \left(1 - \frac{1}{4c_+}\right)\frac{2c_+\sqrt{L}}{4c_+ - 1} = -\frac{\sqrt{L}}{4}.$$

The empirical advantage of a distinguished response depends only on the group's last tokens. It is independent of that response's first $T-1$ tokens and has moments

$$\begin{aligned}
\mathbb{E}[A_c(s_T) \mid \mathscr{F}_{n,k}] &= 0, \\
\mathbb{E}[A_c(s_T)^2 \mid \mathscr{F}_{n,k}] &= \frac{R^2}{4}\left[\left(1 - \frac{1}{|G|}\right)^2 + \frac{|G|-1}{|G|^2}\right] = \frac{R^2(|G|-1)}{4|G|}.
\end{aligned}$$

Consequently $\mathbb{E}[Z_{\mathrm{GRPO}_2} \mid \mathscr{F}_{n,k}] = 0$. Jensen's inequality applied to the full token sum gives

$$\begin{aligned}
\mathbb{E}[\|Z_{\mathrm{GRPO}_2}\|^2 \mid \mathscr{F}_{n,k}] &= \frac{R^2(|G|-1)}{4|G|T^2}\,\mathbb{E}\left[\left(\sum_{t=1}^{T-1} \rho_t^+ g_t^{\theta_{n,k}}\right)^2 \mid \mathscr{F}_{n,k}\right] \\
&\ge \frac{R^2 L(|G|-1)(T-1)^2}{64|G|T^2} \ge \frac{R^2 L(|G|-1)}{256|G|}.
\end{aligned}$$

This proves (46) for every $T \ge 2$.

TIC-GRPO. Keep the same independent reward token and empirical baseline, and set $\theta_{n,k} = \theta_{n,0} = 0$. All likelihood ratios equal one. The first $T-1$ scores are independent, have mean zero, and have second moment L. The empirical advantage remains independent of their sum. Therefore the core has mean zero and

$$\mathbb{E}[\|Z_{\mathrm{TIC}}\|^2 \mid \mathscr{F}_{n,k}] = \frac{R^2 L(|G|-1)(T-1)}{4|G|T^2} \ge \frac{R^2 L(|G|-1)}{8|G|T},$$

which is (47).

Original GRPO. Let all T tokens follow the parameter-dependent policy independently and let $r(s_T) = R$ for the all-zero response and zero otherwise. Use $|G| = 2$ and set

$$\theta_{n,k} = 0, \qquad \theta_{n,0} = -\frac{\log(T^2 - 1)}{2\sqrt{L}}.$$

The old success probability is T^{-2}. Consider the event that the distinguished response has exactly one token 1 and the other response is all zeros. Its probability is

$$\frac{1}{T}(1 - T^{-2})^{2T-1} \ge \frac{1}{T}\left(1 - \frac{2T-1}{T^2}\right) \ge \frac{1}{2T} \qquad (T \ge 4).$$

On this event the distinguished advantage equals $-R/2$. Its token 1 has ratio $T^2/2 > c_+$ and score $\sqrt{L}$, so the negative-advantage branch retains that ratio. Each zero token has score $-\sqrt{L}$ and multiplier between zero and $1/[2(1 - T^{-2})]$. Thus

$$\sum_{t=1}^{T} m_t^{\mathrm{GRPO}} g_t^{\theta_{n,k}} \ge \sqrt{L}\left(\frac{T^2}{2} - \frac{T-1}{2(1-T^{-2})}\right) = \frac{\sqrt{L}\,T^3}{2(T+1)} \ge \frac{\sqrt{L}\,T^2}{4}.$$

It follows that the squared core norm on this event is at least $R^2LT^2/64$, and hence

$$\mathbb{E}[\|Z_{\mathrm{GRPO}}\|^2 \mid \mathscr{F}_{n,k}] \geq \frac{R^2LT}{128}.$$

To bound the mean, use $|A_c| \leq R/2$ for this two-response group, $0 \leq m_t^{\mathrm{GRPO}} \leq \rho_t$, and $|g_t^{\theta_{n,k}}| = \sqrt{L}$. Each old-policy token ratio has expectation one, so

$$|\mathbb{E}[Z_{\mathrm{GRPO}} \mid \mathscr{F}_{n,k}]| \leq \frac{R}{2T} \sum_{t=1}^{T} \mathbb{E}[\rho_t |g_t^{\theta_{n,k}}| \mid \mathscr{F}_{n,k}] = \frac{R\sqrt{L}}{2}.$$

Subtracting the squared mean yields

$$\begin{aligned} &\mathbb{E}\Big[\|Z_{\mathrm{GRPO}} - \mathbb{E}[Z_{\mathrm{GRPO}} \mid \mathscr{F}_{n,k}]\|^2 \mid \mathscr{F}_{n,k}\Big] \\ &\qquad \geq R^2L\left(\frac{T}{128} - \frac{1}{4}\right) \geq \frac{R^2LT}{256} \qquad (T \geq 64). \end{aligned}$$

This proves (48). □

AI use statement

All mathematical proofs and derivations were independently developed and verified by the authors. AI assistance with mathematical verification was limited to generating Lean code and running Lean checks for selected components of the analysis. Generative AI tools (OpenAI ChatGPT and Codex) also assisted with literature discovery and reference verification, manuscript drafting, and LaTeX preparation. The authors take full responsibility for the correctness, originality, and integrity of the paper, including all AI-assisted content.

Ethics statement

This work studies the optimization of reward-based language-model training and evaluates public models on mathematical reasoning and coding benchmarks. The mathematical guarantees concern optimization and update variance; their scope does not include the safety or fairness of deployed language models.

Reproducibility statement

Section 4 specifies the regularity assumptions and conditional sampling protocol. The update rules appear in Sections 2.2 and 3, and the appendices provide complete mathematical proofs and variance constructions. Appendix A describes the models, datasets, training configuration, evaluation metrics, and component ablations, including the normalization and sampling conventions used in the experiments.

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 state the update mechanisms, the response-level variance bounds, the comparison of proved convergence upper bounds, and the empirical scope. Sections 4 and 5 give informal result summaries with formal statements and proofs in Appendices C–F; Section 6 and Appendix A.4 report benchmark results and matched update-magnitude diagnostics.

   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: The theoretical sampling protocol and regularity conditions are stated in Section 4. Section 5 retains the length and sampling terms explicitly. Appendix A.1 gives the empirical scope and training protocol.

   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: The theoretical results state their assumptions, including Assumptions 4.1 and 4.2 and the conditional sampling protocol. Appendix B proves the score and objective lemmas; Appendices C and D give formal convergence statements and proofs, and Appendices E and F give formal response-core upper and lower bounds and their proofs.

   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: [No]

   Justification: The paper discloses the models, training data, evaluation benchmarks, decoding protocols, batch sizes, update schedule, disabled confounders, and compute resources in Section 6 and Appendix A.1. Complete training configurations, random seeds, and evaluation aggregation records are not included.

   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: [No]

   Justification: The paper provides the algorithmic description and experimental settings, but it does not include an open code release or packaged reproduction scripts. The external models, datasets, and framework used in the experiments are cited in Appendix A.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines 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 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: [No]

   Justification: Appendix A.1 specifies the Qwen3 model sizes, hybrid training data, evaluation benchmarks, Avg@1/Avg@32 protocols, sampling parameters, batch and mini-batch sizes, trajectory reuse, disabled training features, and compute configuration. A complete configuration with all training hyperparameters is not included.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

Justification: The paper reports benchmark scores and training curves, while Appendix A.4 reports temporal means and standard deviations of squared update norms across optimizer steps within a run. These describe training dynamics rather than uncertainty across independently trained models; confidence intervals and formal significance tests for the benchmark differences are not reported.

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: [No]

   Justification: Appendix A.1 reports a single H200 GPU for approximately 24 hours for Qwen3-1.7B and two H200 nodes for 48 hours for Qwen3-8B. The number of GPUs per node, memory configuration, and resource usage of the Qwen3-0.6B diagnostics are not specified.

   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?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on RLHF optimization and uses publicly described models, benchmarks, and training corpora. The experiments optimize benchmark rewards using the cited models and datasets.

   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: [No]

    Justification: The paper studies optimization behavior and benchmark performance; it does not include a dedicated discussion of broader societal impacts.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper proposes an optimization algorithm and does not release a new pretrained language model, image generator, scraped dataset, or other high-risk asset.

    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: [No]

    Justification: The paper cites the existing models, datasets, benchmarks, and software framework used in the experiments, including Qwen3, DAPO-Math-17K, AIME, MATH500, Live-CodeBench, and VeRL. The manuscript does not enumerate the license and terms of use for each asset.

    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, the Papers with Code dataset index 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: [N/A]

    Justification: The paper does not introduce or release a new dataset, model checkpoint, or software asset alongside the paper. The new contribution is an algorithmic modification and its analysis.

    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 does not involve crowdsourcing experiments or 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 does not involve crowdsourcing or human-subjects research, so IRB approval or equivalent review 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: [Yes]

    Justification: The AI use statement states that all mathematical proofs and derivations were independently developed and verified by the authors. AI assistance with mathematical verification was limited to generating Lean code and running Lean checks for selected components of the analysis. The statement also discloses assistance from OpenAI ChatGPT and Codex with literature discovery and reference verification, manuscript drafting, and LaTeX preparation, with responsibility for all AI-assisted content retained by the authors. The experimental use of the Qwen3 model family is described in Section 6 and Appendix A.1.

    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.