Beyond Clipping: Signed Logarithmic Smoothing for Policy Optimization

Jihun Yun[1*] **Sungjoon Yoon**[2*] **Beomhan Baek**[3†] **Minhak Song**[4†]
Jongha Jon Ryu[5] **Kwang-Sung Jun**[6]
[1]KRAFTON [2]Boston University [3]University of California, Los Angeles
[4]Stanford University [5]Pennsylvania State University [6]POSTECH GSAI/CSE
jihuny@krafton.com sungjoon@bu.edu kwangsungjun@postech.ac.kr

Abstract

Policy optimization algorithms that use rollouts often optimize ρA where ρ is the importance ratio from the current policy to the rollout policy and A is an advantage estimate. Widely-used algorithms such as PPO and GRPO apply clipping to the ratio ρ as a way to smooth it out, which, depending on ρ, zeros out the gradient. Trading off bias and variance, clipping helps optimization. However, *is clipping the best we can do?* In this paper, we propose an alternative smoothing operator called signed logarithmic smoothing (Signed LS), $\psi_\beta(z) = \mathrm{sgn}(z)\,\beta^{-1}\log(1+\beta|z|)$ and apply it to the ratio times the advantage ρA rather than the ratio ρ alone. Building on Signed LS, we introduce KRAFT, a GRPO-style objective that replaces clipping with Signed LS. In an off-policy contextual bandit model, we establish theoretical results characterizing high-probability approximation error bounds of smoothing operators. Our analysis shows that, depending on the smoothing level, clipping and Signed LS show different tradeoffs. In particular, with best tuning parameters, we show that Signed LS achieves a better bound than clipping. In contextual bandit experiments, KRAFT outperforms clipping baselines and shows lower hyperparameter sensitivity. We also validate our approach by fine-tuning three LLMs for mathematical reasoning. KRAFT matches or exceeds GRPO's average benchmark accuracy, with comparable or lower reference KL and higher policy entropy at the end of training.

1 Introduction

Policy optimization typically updates a target policy using rollouts collected by a behavior policy [1–4]. Importance ratios correct this policy mismatch and are standard in off-policy contextual-bandit settings [5–7]. Policy optimization in language models has the same structure: prompts are contexts and responses provide rewards, verifier scores, or advantages. PPO [4] was a practical breakthrough in policy optimization. PPO uses a clipped importance-ratio surrogate to restrict policy updates. In the sign-dependent clipped tails, the surrogate is constant in the ratio and its gradient vanishes. This clipped-ratio design carries over to GRPO [8], which uses group-relative reward normalization. GRPO has shown strong empirical performance in reasoning RL [9]. Such clipping design is widely used by DAPO [10] and Dr. GRPO [11]. Also, GDPO [12] and REINFORCE++ [13] still use clipping, as do many other recent policy optimization methods. In this context, we ask:

Is clipping the best we can do?

[*]Equal contribution
[†]Work done during internships at KRAFTON.

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

We answer this question by proposing signed logarithmic smoothing (Signed LS), a smooth alternative to ratio clipping. Signed LS is a signed extension of the logarithmic smoothing operator, which first appeared in Audibert and Bubeck [14] (see $v_{i,t}$ therein), to our knowledge, and was later used in off-policy contextual bandits [15]. Signed LS applies $\rho A \mapsto \psi_\beta(\rho A)$ in policy optimization where

$$\psi_\beta(z) := \operatorname{sgn}(z)\frac{1}{\beta}\log(1+\beta|z|).$$

The logarithmic form preserves a statistical connection to the original importance-weighted objective and acts on ρA rather than on ρ alone. While the clipped surrogate is flat in sign-dependent tails, Signed LS continues to respond to the full weighted update while damping it smoothly. We also introduce **KRAFT**, a GRPO-style method that substitutes the clipped per-token surrogate with Signed LS.

For theoretical justification, we consider off-policy contextual bandits. The main analysis first compares the deviation bound of Signed LS with that of hard and soft clipping. Clipping-style surrogates introduce explicit ratio-dependent bias and threshold-dependent concentration terms, while Signed LS smooths the weighted feedback $Y_t = \rho_t R_t$ and pays a smooth second-order penalty. We prove surrogate fidelity for Signed LS. The corresponding one-sided bounds transfer surrogate regret to the original importance-weighted regret against a fixed comparator policy.

Contributions. We summarize our contributions as follows.

- **Smooth the update, not just the ratio.** Motivated by the zero-gradient regions by clipped-ratio surrogates, we propose Signed LS, which applies $\psi_\beta(\cdot)$ to the full weighted update ρA. Signed LS damps large updates without clipped flat tails. Our surrogate approaches ρA as $\beta \to 0^+$ and bounds the policy-gradient coefficient by $1/\beta$. Built on this surrogate, we propose **KRAFT**, a GRPO-style objective that replaces the clipped per-token surrogate with Signed LS.
- **Fidelity and regret transfer without ratio-tail clipping bias.** We prove surrogate fidelity for Signed LS in off-policy contextual bandits using exact MGF identities and Ville's inequality. The bound holds over the training horizon for $Y_t = \rho_t R_t$ and controls the deviation from the cumulative importance-weighted conditional mean by a Signed-LS second-order term. The corresponding one-sided deviation bounds yield regret transfer to the original importance-weighted regret against a fixed comparator policy. We also contrast the resulting bound with hard and soft clipping, whose decompositions contain ratio-dependent bias and threshold-dependent concentration terms. With tuned parameters, Signed LS improves bounds over the hard/soft clipping.
- **Higher accuracy with a tendency toward lower KL and higher entropy.** In our contextual bandit experiments, KRAFT achieves the highest mean reward across the epochs and the lowest overall hyperparameter sensitivity. We also evaluate our method in RLVR fine-tuning on standard math reasoning problems. Overall, KRAFT improves GRPO in terms of not only accuracy (higher) but also reference KL (lower) and entropy (higher); see Figure 2.

2 Preliminaries

Notations. We denote the response length by L and the token position by ℓ. For a group of G responses generated from the same prompt, L_i denotes the length of the i-th response. In the contextual-bandit analysis, t denotes the current round and T the total number of rounds. We write μ_t and π_t for the behavior and target policies at round t, respectively.

RLHF and Policy Optimization. Reinforcement learning from human feedback (RLHF) [16, 17] trains a policy π_θ against a learned reward model with KL regularization toward a reference policy π_{ref}. Policy optimization uses stochastic policy-gradient updates $\mathbb{E}_{x\sim\mathcal{D},a\sim\pi_\theta(\cdot|x)}[\nabla_\theta \log \pi_\theta(a|x)A(x,a)]$ [2], where $A(x,a)$ is an advantage estimate. In practice, rollouts collected under a behavior policy $\pi_{\theta_{\text{old}}}$ are reused across gradient steps. The importance ratio $\rho(\theta) = \pi_\theta(a|x)/\pi_{\theta_{\text{old}}}(a|x)$ corrects the policy mismatch. Large ratios can destabilize training. Trust-region methods address this instability by constraining the policy update [3]. PPO replaces the constraint with a clipped surrogate [4].

Proximal Policy Optimization (PPO). PPO [4] replaces the trust-region constraint with a clipped surrogate to limit policy change per update. For a prompt $x \sim \mathcal{D}$, let $y = (y_1, \ldots, y_L)$ be a response

of length L and $\mathsf{clip}_\varepsilon(\rho) = \mathsf{clip}(\rho, 1-\varepsilon, 1+\varepsilon)$.

$$\mathcal{J}^{\mathsf{PPO}}(\theta) = \mathbb{E}_{x,y}\left[\frac{1}{L}\sum_{\ell=1}^{L}\min\Big(\rho_\ell(\theta)A_\ell, \mathsf{clip}_\varepsilon\big(\rho_\ell(\theta)\big)A_\ell\Big)\right], \tag{1}$$

where $\rho_\ell(\theta) = \pi_\theta(y_\ell|x, y_{<\ell})/\pi_{\theta_{\text{old}}}(y_\ell|x, y_{<\ell})$ is the token importance ratio. The estimated advantage is A_ℓ and the clipping radius is ε. PPO maximizes the objective in Equation (1).

Group Relative Policy Optimization (GRPO). GRPO [8] removes the critic by normalizing the rewards of responses generated from the same prompt. For prompt x, let G be the group size and $y_i = (y_{i,1}, \ldots, y_{i,L_i})$ be the i-th generated response. The per-token objective is

$$\mathcal{J}^{\mathsf{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^{G}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{\ell=1}^{L_i}\min\Big(\rho_{i,\ell}(\theta)A_i,\ \mathsf{clip}_\varepsilon(\rho_{i,\ell}(\theta))A_i\Big) - \alpha\mathrm{KL}\left[\pi_\theta\|\pi_{\mathsf{ref}}\right]\right], \tag{2}$$

where $\rho_{i,\ell}(\theta) = \pi_\theta(y_{i,\ell}|x, y_{i,<\ell})/\pi_{\theta_{\text{old}}}(y_{i,\ell}|x, y_{i,<\ell})$ is the per-token importance ratio. The response-level advantage is $A_i = (r(x, y_i) - \bar{r})/\sigma_r$. The group mean is $\bar{r} = G^{-1}\sum_i r(x, y_i)$ and the group standard deviation is σ_r. All advantages are set to zero when the group rewards are identical. The coefficient $\alpha > 0$ controls KL regularization toward π_{ref}. PPO incorporates a KL penalty into the reward signal and estimates token-level advantages using a value function. GRPO adds the KL penalty directly to the objective. GRPO shares the response-level advantage A_i across all tokens in y_i and therefore does not need a critic. Both objectives maximize the clipped surrogate as a trust-region approximation.

Online Contextual Bandits and Regret. The following contextual-bandit model underlies our regret bounds and surrogate fidelity results. We denote the history through the end of round t by $\mathcal{F}_t$, so $\mathcal{F}_{t-1}$ contains the information available before that round begins. Let π_t be the target policy and μ_t the behavior policy at round t. After drawing $x_t \sim \mathcal{D}$, we sample $a_t \sim \mu_t(\cdot|x_t)$ and observe a reward $R_t \in [-R_{\mathsf{max}}, R_{\mathsf{max}}]$. We assume a stationary reward $r(x, a)$ with $\mathbb{E}[R_t \mid \mathcal{F}_{t-1}, x_t, a_t] = r(x_t, a_t)$, and we define the value of a policy as $J(\pi) \triangleq \mathbb{E}_{x\sim\mathcal{D}, a\sim\pi(\cdot|x)}[r(x, a)]$. We assume that an optimal policy $\pi^* \in \arg\max_\pi J(\pi)$ exists. The regret is

$$\mathcal{R}_T = \sum_{\tau=1}^{T}\big[J(\pi^*) - J(\pi_\tau)\big]. \tag{3}$$

Under support coverage, letting $\rho_t = \pi_t(a_t|x_t)/\mu_t(a_t|x_t)$ be the importance ratio, the importance-weighted reward $\widetilde{R}_t = \rho_t R_t$ has $\mathbb{E}[\widetilde{R}_t \mid \mathcal{F}_{t-1}, x_t] = \mathbb{E}_{a\sim\pi_t(\cdot|x_t)}[r(x_t, a)]$. Thus, $\widetilde{R}_t$ is an unbiased per-step estimate of the target-policy value conditional on the past. Our results allow π_t and μ_t to be $\mathcal{F}_{t-1}$-measurable. Multi-epoch reuse requires further assumptions such as uniform convergence or stability.

3 Beyond Clipping: Signed Logarithmic Smoothing

PPO/GRPO clipping acts on ρ alone and introduces a layer of nonsmoothness while being flat in sign-dependent tails. We replace ρu by a smooth signed logarithmic transform that preserves the sign of u and damps large $|\rho u|$. Existing logarithmic smoothing for off-policy contextual bandits is one-sided [15] and can be undefined for signed advantages when $\rho u \le -1/\beta$. For $\beta > 0$, Signed LS uses

$$\psi_\beta(z) := \mathrm{sgn}(z)\frac{1}{\beta}\log(1 + \beta|z|).$$

Given $\rho > 0$ and an advantage $u \in \mathbb{R}$, the Signed LS surrogate is

$$\mathsf{SLS}_\beta(\rho, u) := \psi_\beta(\rho u).$$

The map ψ_β is odd and C^1 and has derivative

$$\psi'_\beta(z) = \frac{1}{1 + \beta|z|}.$$

Damping becomes substantial around $|\rho u| \approx 1/\beta$. With u fixed, the policy-update gradient is

$$\nabla_\theta \psi_\beta(\rho_\theta u) = \frac{\rho_\theta u}{1+\beta|\rho_\theta u|}\nabla_\theta \log \pi_\theta(a|x).$$

The coefficient $\rho_\theta u/(1+\beta|\rho_\theta u|)$ on $\nabla_\theta \log \pi_\theta(a|x)$ nearly matches $\rho_\theta u$ for small $|\rho_\theta u|$ and is bounded in magnitude by $1/\beta$ for every $\rho_\theta u$. The coefficient bound yields $\|\nabla_\theta \psi_\beta(\rho_\theta u)\| \le C/\beta$ whenever the log-probability gradient satisfies $\|\nabla_\theta \log \pi_\theta(a \mid x)\| \le C$. In stark contrast, PPO-style clipping makes the clipped surrogate flat in sign-dependent tails: above the upper threshold for $u > 0$ and below the lower threshold for $u < 0$. These introduce a layer of nonsmoothness and induce zero gradients when the clipping is active, which could slow down the optimization process.

Exponential-moment structure. Signed LS provides a smooth alternative to clipping and preserves an explicit statistical connection to the original importance-weighted feedback. Direct algebra with $g(z) := z^2/(1+\beta|z|)$ gives

$$e^{\beta\psi_\beta(z)} = 1+\beta z+\mathbf{1}\{z<0\}\beta^2 g(z), \quad e^{-\beta\psi_\beta(z)} = 1-\beta z+\mathbf{1}\{z\ge 0\}\beta^2 g(z).$$

Thus, the exponential transform keeps the first-order component equal to the original input z while the price of damping appears as an explicit nonnegative second-order correction. Clipping truncates terms beyond a threshold, which can change policy ordering. Section 4 applies these identities to $Y_t = \rho_t R_t$ and obtains surrogate fidelity via Ville's inequality.

KRAFT. Based on Signed LS, KRAFT incorporates GRPO's group-relative objective by replacing the clipped per-token surrogate in Equation (2) with $\psi_\beta(\rho_{i,\ell}A_i)$. Under the notation of Equation (2), the KRAFT objective is

$$\mathcal{J}^{\mathsf{KRAFT}}(\theta) = \mathbb{E}_{x,\{y_i\}_{i=1}^G}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{\ell=1}^{L_i}\psi_\beta(\rho_{i,\ell}(\theta)A_i) - \alpha \mathrm{KL}[\pi_\theta\|\pi_{\mathsf{ref}}]\right].$$

For finite β, large $|\rho_{i,\ell}A_i|$ are damped rather than clipped.

4 Theoretical Analysis

In this section, we justify Signed LS as a statistically sound surrogate for the original importance-weighted objective. We consider an off-policy contextual bandit in which behavior policies generate data and importance ratios evaluate target policies. This setting isolates the statistical issue behind rollout reuse. We first compare the deviation structure of Signed LS with hard and soft clipping, and then use the Signed LS guarantee to transfer surrogate regret to the original importance-weighted objective. For binary rewards, we also compare the expectations of the GRPO and KRAFT stochastic gradient estimators.

4.1 Deviation Bounds for Clipping and Signed LS

We analyze Signed LS in an adaptive off-policy setting with predictable policies: π_t, μ_t may depend on the past but are fixed before the current round is observed. We first bound the deviation between the cumulative surrogate and the cumulative importance-weighted conditional mean (Theorem 1) and compare this structure with hard and soft clipping. Section 4.2 then turns this fidelity guarantee into regret transfer for the original importance-weighted objective.

Setup. Fix horizon $T \in \mathbb{N}$ and smoothing parameter $\beta > 0$. The action space $\mathcal{A}$ is countable. We denote the history through the end of round t by $\mathcal{F}_t$ and write $\mathbb{E}_{t-1}[\cdot]$ for conditional expectation given $\mathcal{F}_{t-1}$. At round t, the behavior μ_t and target π_t are $\mathcal{F}_{t-1}$-measurable. We then observe x_t and sample $a_t \sim \mu_t(\cdot|x_t)$ before observing the reward R_t. We define $\rho_t := \pi_t(a_t|x_t)/\mu_t(a_t|x_t)$ on the behavior support with $\rho_t := 0$ otherwise, and we set $Y_t := \rho_t R_t$. We use the Signed LS map $\psi_\beta(z) := \mathrm{sgn}(z)\frac{1}{\beta}\log(1+\beta|z|)$ from Section 3.

For each $t \le T$, we define $v_t := \mathbb{E}_{t-1}[Y_t]$, $g(y) := \frac{y^2}{1+\beta|y|}$ and

$$S_t := \sum_{\tau=1}^{t} \psi_\beta(Y_\tau), \qquad V_t := \sum_{\tau=1}^{t} v_\tau,$$

$$Q_t := \sum_{\tau=1}^{t} \mathbb{E}_{\tau-1}[g(Y_\tau)], \qquad Q_t^{\pm} := \sum_{\tau=1}^{t} \mathbb{E}_{\tau-1}\big[\mathbf{1}\{\pm Y_\tau > 0\} g(Y_\tau)\big].$$

Note that $Q_t = Q_t^- + Q_t^+$. Suppose also that $R_t = r(x_t, a_t) + \xi_t$ for a policy-independent reward function r and noise satisfying $\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1} \vee \sigma(x_t, a_t)] = 0$ a.s. When these conditional expectations are finite, we have

$$v_t = \mathbb{E}_{t-1}\left[\sum_a \pi_t(a|x_t) r(x_t, a)\right],$$

so v_t is the conditional value of the target policy at round t. The derivation is given in Appendix A.1.

Assumption 1 (Support coverage)**.** *For every round $t \in \{1, \dots, T\}$, $\pi_t(\cdot|x_t)$ is absolutely continuous with respect to $\mu_t(\cdot|x_t)$ almost surely. Equivalently, for every action $a \in \mathcal{A}$, $\pi_t(a|x_t) > 0 \Rightarrow \mu_t(a|x_t) > 0$ a.s.*

Assumption 2 (Bounded rewards)**.** *We have $|R_t| \le 1$ almost surely for every round t.*

Assumption 1 is the standard coverage, or absolute-continuity, condition for importance-sampling-based off-policy estimation [18–20]. Together with Assumption 2, it directly implies $Q_T \le T/\beta$ a.s. even when the importance ratios are unbounded.

4.1.1 Signed LS

We bound how far the Signed LS surrogate S_t can deviate from the cumulative target value V_t over the training horizon.

Theorem 1 (Surrogate fidelity)**.** *Suppose Assumptions 1 and 2 hold. For any $\delta \in (0, 1)$, the following bound holds simultaneously for all $1 \le t \le T$ with probability at least $1 - \delta$:*

$$\big|S_t - V_t\big| \le \beta Q_t + \frac{1}{\beta} \log\left(\frac{2}{\delta}\right).$$

Theorem 1 says that Signed LS remains a faithful proxy for cumulative importance-weighted feedback uniformly over the training horizon. Smoothing may compress large values, but the deviation is controlled by the second-order quantity Q_t and the scale β, in line with recent second-order analyses for offline contextual bandits [21]. We give the proof and a bound for heavy-tailed ratios in Appendix A.2 (Proposition A.3).

4.1.2 Hard and Soft Clipping

Both clipping bounds include a term for the cumulative bias introduced by shrinking the importance weights.

Hard clipping. For a clipping threshold $\kappa > 0$, we define the cumulative clipped surrogate by

$$S_t^{\mathsf{clip}}(\kappa) := \sum_{\tau=1}^{t} \min\{\rho_\tau, \kappa\} R_\tau,$$

so that $\widehat{v}_n^{\mathsf{clip}}(\kappa) = \frac{1}{n} S_n^{\mathsf{clip}}(\kappa)$. The conditional clipping bias at round t is

$$b_t(\kappa) := v_t - \mathbb{E}_{t-1}\left[\min\{\rho_t, \kappa\} R_t\right] = \mathbb{E}_{t-1}\left[(\rho_t - \kappa)\mathbf{1}\{\rho_t \ge \kappa\} R_t\right],$$

and we write $B_t^{\mathsf{clip}}(\kappa) := \sum_{\tau=1}^{t} b_\tau(\kappa)$ for its cumulative version. The cumulative conditional variance is

$$\Sigma_t^2(\kappa) := \sum_{\tau=1}^{t} \mathrm{Var}_{\tau-1}\left(\min\{\rho_\tau, \kappa\} R_\tau\right).$$

Theorem 2 (Surrogate fidelity: hard clipping)**.** *Suppose Assumptions 1 and 2 hold. Let $\kappa > 0$ and $\lambda \in (0, 3/(2\kappa))$. For any $\delta \in (0, 1)$, the following inequality holds simultaneously for all*

$t \in \{1, \dots, T\}$ *with probability at least* $1 - \delta$*:*

$$\left|S_t^{\mathsf{clip}}(\kappa) - V_t\right| \le |B_t^{\mathsf{clip}}(\kappa)| + \frac{\lambda \Sigma_t^2(\kappa)}{2(1 - 2\kappa\lambda/3)} + \frac{1}{\lambda} \log \frac{2}{\delta}.$$

The proof is deferred to Appendix A.3.

Soft clipping. One can also consider a soft version of clipping:

$$S_t^{\mathsf{sclip}}(\kappa) := \sum_{\tau=1}^{t} \frac{\rho_\tau R_\tau}{1 + \rho_\tau/\kappa}.$$

This has the following conditional bias and the cumulative conditional variance:

$$b_t{}^{\mathsf{sclip}}(\kappa) := \mathbb{E}_{t-1}\left[\frac{{\rho_t}^2 R_t}{\kappa + \rho_t}\right], \quad (\Sigma_t^{\mathsf{sclip}}(\kappa))^2 := \sum_{\tau=1}^{t} \mathrm{Var}_{\tau-1}\left(\frac{\rho_\tau R_\tau}{1 + \rho_\tau/\kappa}\right).$$

We write $B_t^{\mathsf{sclip}}(\kappa) := \sum_{\tau=1}^{t} b_\tau^{\mathsf{sclip}}(\kappa)$ for the cumulative bias.

Theorem 3 (Surrogate fidelity: soft clipping)**.** *Suppose Assumptions 1 and 2 hold. Let* $\kappa > 0$ *and* $\lambda \in (0, 3/(2\kappa))$*. For any* $\delta \in (0, 1)$*, the following inequality holds simultaneously for all* $t \in \{1, \dots, T\}$ *with probability at least* $1 - \delta$*:*

$$\left|S_t^{\mathsf{sclip}}(\kappa) - V_t\right| \le |B_t^{\mathsf{sclip}}(\kappa)| + \frac{\lambda (\Sigma_t^{\mathsf{sclip}}(\kappa))^2}{2(1 - 2\kappa\lambda/3)} + \frac{1}{\lambda} \log \frac{2}{\delta}.$$

Comparison at a fixed sample size. Consider a fixed pair of policies under support coverage and n i.i.d. samples with $0 \le R \le 1$. Define $m_2 = \mathbb{E}[(\rho R)^2]$ and assume $h_2 = \mathbb{E}[\rho^2 R] < \infty$. At fixed confidence, the tuned bounds scale as $\sqrt{m_2/n}$ for Signed LS and $\sqrt{h_2/n}$ for soft clipping (Proposition A.4). Clipping determines attenuation from the ratio alone, whereas Signed LS adapts it to the weighted reward. Since $R^2 \le R$, we have $m_2 \le h_2$, with equality for binary rewards. A numerical comparison on a two-context instance shows the same behavior (Appendix C).

4.2 Regret Transfer for Signed LS

We now extend our fidelity result to a regret comparison against a fixed comparator policy. For any fixed policy $\widetilde{\pi}$, we let $S_T(\widetilde{\pi})$, $V_T(\widetilde{\pi})$, and $Q_T^+(\widetilde{\pi})$ denote the corresponding quantities obtained by evaluating $\widetilde{\pi}$ as a static target policy under the same behavior policies. Suppose this static policy also satisfies Assumption 1. We define $R_T^{\mathsf{surr}}(\widetilde{\pi}) := S_T(\widetilde{\pi}) - S_T$ and $R_T^{\mathsf{IW}}(\widetilde{\pi}) := V_T(\widetilde{\pi}) - V_T$.

Corollary 1 (Regret transfer)**.** *Suppose Assumption 2 holds and Assumption 1 holds for both the learner and* $\widetilde{\pi}$*. For any* $\delta \in (0, 1)$*, the following holds with probability at least* $1 - \delta$*:*

$$R_T^{\mathsf{IW}}(\widetilde{\pi}) \le R_T^{\mathsf{surr}}(\widetilde{\pi}) + \beta(Q_T^+(\widetilde{\pi}) + Q_T^-) + \frac{2}{\beta} \log\left(\frac{2}{\delta}\right).$$

Corollary 1 yields a sublinear upper bound on regret for the original objective when the surrogate regret and the additional terms in the bound are all $o(T)$.

Connection to RLVR. Theorem 1 allows policy changes between rounds. Under the behavior policy, each token ratio has mean one given the prompt and preceding tokens. In GRPO, every token in response i shares the same advantage A_i, which is computed from the rewards of all responses in the group. This dependence on the full group prevents a direct token-by-token application of Theorem 1. Reusing rollouts across epochs requires controlling policy changes. The KL penalty cancels in the difference between the smoothed and importance-weighted objectives. For a complete response, the product of token ratios gives the importance ratio for a single bandit action (Appendix A.2, Corollary A.1).

4.3 Effect on SGD Updates

We now turn from surrogate fidelity to how smoothing changes the policy update. Even for softmax policy gradient with linear function approximation, global convergence remains an active question [22]. Full GRPO convergence remains open, so we compare GRPO and KRAFT under local SGD.

For this purpose, we fix a prompt x with a finite action set $\mathcal{A}$ and consider binary rewards. The expected reward $J_x(\theta) = \mathbb{E}_{\pi_\theta}[R \mid x]$ is the probability of success. To model a GRPO update, we draw $G \geq 2$ independent action–reward pairs from the current policy and normalize the rewards in the sampled group to obtain advantages A_i. At unit importance ratio, KRAFT replaces each A_i in the gradient estimator with $A_i/(1 + \beta|A_i|)$ for $\beta > 0$.

A group yields a nonzero advantage only if it contains both successes and failures. We therefore consider $0 < J_x(\theta) < 1$. In this case, the expected GRPO update is a positive multiple of $\nabla J_x(\theta)$. At the same θ, the expected KRAFT update is $c \in (0, 1)$ times the expected GRPO update. We compare the two methods after N steps from $\theta_0^{\mathsf{GRPO}} = \theta_0^{\mathsf{KRAFT}} = \theta_0$. Each method samples from its own policy at each step using a constant learning rate (Appendix A.4). We assume that $\pi_\theta(a \mid x) > 0$ is C^3 in a neighborhood of θ_0 and that the reward distribution given (x, a) does not depend on θ.

Proposition 1 (Learning Rate Comparison)**.** *Suppose $\nabla J_x(\theta_0) \neq 0$ and let $N \geq 1$. For every sufficiently small $\Delta > 0$, there are locally unique positive learning rates $\eta(\Delta)$ for GRPO and $\widetilde{\eta}(\Delta)$ for KRAFT satisfying*

$$\mathbb{E}[J_x(\theta_N^{\mathsf{GRPO}}(\eta))] = \mathbb{E}[J_x(\theta_N^{\mathsf{KRAFT}}(\widetilde{\eta}))] = J_x(\theta_0) + \Delta.$$

These rates converge to zero as $\Delta \downarrow 0$ and satisfy

$$\frac{\widetilde{\eta}(\Delta)}{\eta(\Delta)} = \frac{1}{c} + O(\Delta) > 1,$$

where c is evaluated at θ_0.

KRAFT therefore needs a higher learning rate to compensate for its smaller expected gradient in this local SGD setting. Whether this carries over to Adam remains open.

5 Experiments

We evaluate KRAFT in two settings: a contextual bandit and an RLVR study on mathematical reasoning. The primary baseline is GRPO [8], the clipped-ratio method that KRAFT replaces. We measure performance by reward in the bandit study and by task accuracy in the RLVR study. We also track reference KL and policy entropy during RLVR training. The bandit study includes clipping variants and SAPO-sym [23] as additional baselines. Ratio-only LS isolates the effect of smoothing ρA rather than ρ alone.

5.1 Contextual Bandits

Bandit Setup. The bandit study isolates optimization behavior: the reward function is known and evaluation is noise-free. A bilinear function of the context and action features is perturbed by Gaussian noise and passed through a bounded nonlinearity. The resulting rewards lie in $(-1, 1)$. We train a softmax policy off-policy using actions sampled from the previous policy iterate. Evaluation uses expected noiseless reward on 1,000 held-out contexts. Each update draws a small group of actions per context and normalizes their rewards to form advantages, as in GRPO. The batch is reused for E gradient steps before new actions are drawn. Thus E is the number of PPO epochs.

We compare the unclipped importance-weighted baseline, PPO-style two-sided clipping, one-sided clipping, SAPO-sym, Ratio-only LS, and KRAFT. Ratio-only LS is an ablation defined as $\mathcal{L}^{\mathsf{ratio}} = -\beta^{-1}\log(1 + \beta\rho)A$, which smooths ρ but leaves A outside the logarithm. SAPO [23] replaces clipping with a smooth sigmoid surrogate using separate hyperparameters τ_+ and τ_-. We use a symmetric variant called SAPO-sym with gate $\sigma(\tau(\rho - 1)) \cdot 4/\tau$. SAPO-sym uses a single parameter τ, so each compared surrogate has one tuning parameter. See Appendix B for setup and hyperparameters.

Robustness to PPO Epochs. PPO uses clipping to constrain ratio growth when a batch is reused across gradient steps. Table 1 reports test reward as the number of PPO epochs E varies from 1 to 5. KRAFT outperforms the Baseline, both clipping variants, SAPO-sym, and Ratio-only LS for $E \in \{2, 3, 4, 5\}$ ($p < 0.05$ under a paired t test). KRAFT has the highest mean reward at every compared epoch setting. Its largest reported mean is 0.7460 at $E = 2$. Ratio-only LS peaks at 0.7318 and SAPO-sym at 0.7303, both at $E = 3$. The Baseline declines monotonically from 0.7230 to

Table 1: Test reward versus the number of PPO epochs E (mean $\pm$ 1 SE over 200 seeds). Each method uses its best (η, λ) selected at the final step. **Bold** marks the best value in each row. For every $E \geq 2$, the gap between KRAFT and each other method is significant at the 0.05 level under a paired t test.

E	Baseline	PPO Clipping	Clipping	SAPO-sym	Ratio-only LS	KRAFT
1	0.7230 $\pm$0.0038	—	—	—	—	—
2	0.7058 $\pm$0.0035	0.7271 $\pm$0.0039	0.7314 $\pm$0.0041	0.7299 $\pm$0.0031	0.7284 $\pm$0.0032	**0.7460 $\pm$0.0033**
3	0.6826 $\pm$0.0039	0.7222 $\pm$0.0034	0.7239 $\pm$0.0041	0.7303 $\pm$0.0036	0.7318 $\pm$0.0036	**0.7426 $\pm$0.0036**
4	0.6621 $\pm$0.0067	0.7173 $\pm$0.0036	0.7206 $\pm$0.0036	0.7287 $\pm$0.0032	0.7298 $\pm$0.0038	**0.7429 $\pm$0.0034**
5	0.6591 $\pm$0.0047	0.7169 $\pm$0.0037	0.7174 $\pm$0.0038	0.7290 $\pm$0.0033	0.7288 $\pm$0.0037	**0.7447 $\pm$0.0036**

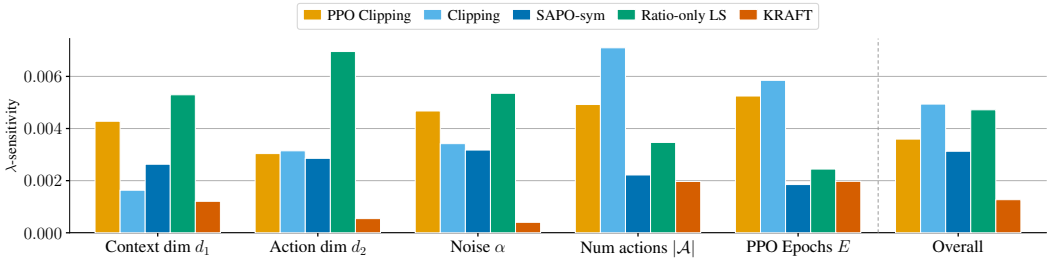


Figure 1: λ-sensitivity (*lower is better*): Per-setting tuned score minus cross-setting score with λ fixed to λ^* and η tuned per setting. KRAFT has the smallest overall value and leads in four of five sweeps.

0.6591. Both clipping variants peak at $E = 2$ and then decline. Smoothing the full product ρA yields higher mean rewards than Ratio-only LS and SAPO-sym at each compared epoch setting.

Tuning Sensitivity. We write λ for each method's primary surrogate hyperparameter. The parameter λ is ε for the clipping methods, τ for SAPO-sym, and β for Ratio-only LS and KRAFT. In practice, λ is often fixed across settings. Sensitivity measures the resulting performance cost. Figure 1 shows λ-sensitivity following a variant of the protocol of Adkins et al. [24]. For each method, we select λ^* to maximize mean reward over all settings. We then fix $\lambda = \lambda^*$ and retune η for each setting. For each sweep, we report the average gap from the best performance obtained by tuning both (λ, η) for each setting. This isolates sensitivity to λ. The overall bar aggregates this gap across all settings. KRAFT has the smallest overall sensitivity and leads in four of five sweeps. This suggests that a single β transfers well across settings. Appendix B.4 gives the details and numerical values in Table 6.

Additional Results. We additionally provide learning curves and robustness sweeps across context dimension, action dimension, reward noise, and number of actions. KRAFT outperforms all baselines at every update step and matches or beats them across all sweep settings, with the margin over clipping-based methods widening as context dimension and noise increase. We also plot reward against policy entropy for KRAFT and PPO Clipping. KRAFT generally achieves higher mean reward than PPO Clipping at comparable intermediate policy entropy. Details are in Appendix B.4.

5.2 LLM Experiments

We evaluate KRAFT in an RLVR setting for mathematical reasoning. GRPO-style clipping is the dominant baseline. This context motivates the following question:

> **Evaluation Question.** Can KRAFT maintain GRPO's task performance while producing more favorable training behavior?

Setup. We follow the RLVR math-reasoning protocol of Xiong et al. [25]. We fine-tune three models on NuminaMath [26]: LLaMA-3.2-3B-Instruct [27], Phi-4-Mini-Instruct [28], and Qwen3-4B-Base [29]. The batch and mini-batch sizes are 1024 and 512, respectively. We use a group size

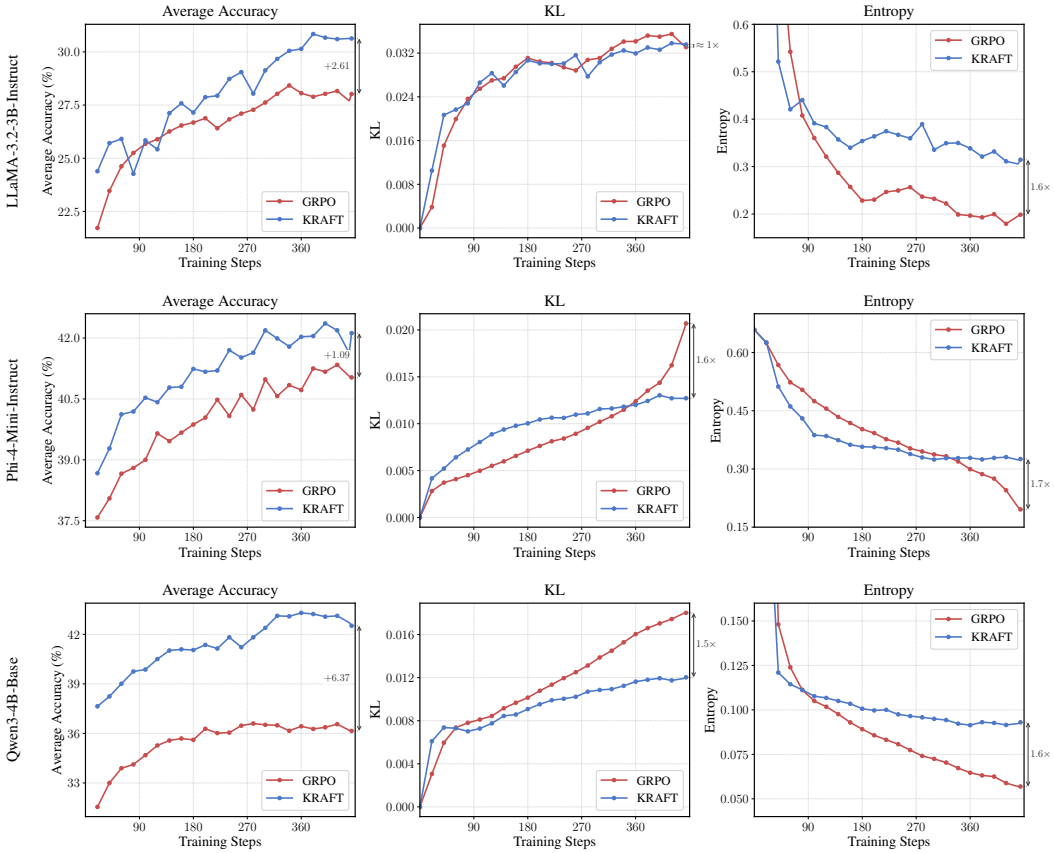


Figure 2: Training curves on NuminaMath for LLaMA-3.2-3B-Instruct, Phi-4-Mini-Instruct, and Qwen3-4B-Base (top to bottom). Columns show the three-benchmark average accuracy, reference KL, and policy entropy. Accuracy annotations give KRAFT's gain over GRPO in percentage points at the final plotted checkpoint. At that checkpoint, KL ratios compare GRPO to KRAFT. Entropy ratios compare KRAFT to GRPO.

of $G = 4$, a maximum response length of 4096 tokens, and a KL coefficient of $\alpha = 10^{-3}$. We evaluate every 20 global steps using average@16 at temperature 1.0 on MATH-500 [30], Minerva Math [31], and OlympiadBench [32]. We use the three-benchmark average as the main aggregate metric. Following Xiong et al. [25], we exclude AIME since it contains only 30 problems, which makes the results noisy. We compare standard GRPO ($\varepsilon = 0.2$ clipping) against KRAFT with $\beta = 1.0$. We choose the learning rates empirically and use 3×10^{-6} for GRPO and 5×10^{-6} for KRAFT. The larger rate for KRAFT is consistent with Proposition 1.

Main Results. Figure 2 compares average accuracy, reference KL, and policy entropy over the course of training. KRAFT achieves higher average accuracy than GRPO across the three models, with a clear advantage on LLaMA-3.2-3B-Instruct and Qwen3-4B-Base and a smaller but sustained improvement on Phi-4-Mini-Instruct. As training progresses, KRAFT shows a tendency toward lower reference KL and higher policy entropy relative to GRPO. This distinction is especially clear later in training on Phi-4-Mini-Instruct, where GRPO's KL rises sharply and its entropy drops, while KRAFT follows a more gradual trajectory. The higher entropy suggests a broader distribution over possible outputs, while the KL behavior indicates more moderate divergence from the reference policy.

Effect of the Smoothing Parameter. We examine the effect of β on LLaMA-3.2-3B-Instruct using the values 0.1, 1, and 10. In this ablation, all methods use a learning rate of 10^{-6} and a mini-batch size of 256, with the other settings unchanged. KRAFT achieves similar average accuracy at $\beta = 0.1$ and 1, with lower reference KL and higher policy entropy than GRPO (Table 2). Increasing β to 10

Table 2: Effect of β on LLaMA-3.2-3B-Instruct. We present average@16 accuracy (%) for each benchmark and the unweighted mean (Avg.), with the highest value in each accuracy column in bold.

Method	MATH-500	Minerva	Olympiad	Avg.	KL	Entropy
GRPO	42.3	14.4	12.5	23.1	0.0187	1.297
KRAFT ($\beta = 0.1$)	**42.7**	**16.0**	**13.1**	**23.9**	0.0157	1.539
KRAFT ($\beta = 1$)	42.4	15.9	**13.1**	23.8	0.0138	1.584
KRAFT ($\beta = 10$)	36.8	13.1	9.1	19.4	0.0038	2.720

further reduces KL and raises entropy but substantially lowers accuracy. These results support $\beta = 1$ as a practical choice in this setting.

In additional ablations, we compare KRAFT with GRPO at the same learning rate and test whether smoothing ρA improves on smoothing ρ alone. We also compare KRAFT with SAPO on LLaMA. To evaluate performance beyond mathematical reasoning, we test the LLaMA checkpoints on AlpacaEval 2.0 and five general benchmarks. Detailed results are in Appendix D.

6 Related Work

RLHF, RLVR and Clipped Policy Optimization. RLHF aligns language models with human preferences [16, 33, 17, 34]. PPO [4] is a common optimizer. Its critic-free variant GRPO [8] is used in reasoning fine-tuning [9]. Recent RLVR optimizers extend GRPO in several directions. DAPO [10] decouples the lower and upper clipping ranges. GDPO [12] and Dr. GRPO [11] address normalization in multi-reward and group-relative settings. REINFORCE++ [13] uses global advantage normalization in place of group-relative normalization. These variants modify clipping ranges or normalization while using the clipped objective. KRAFT replaces the clipped per-token surrogate with Signed LS applied to ρA.

Smooth or Clipping-Free LLM Policy Optimization. A separate line of work replaces the clip function outright. CFPO [35] uses a quadratic penalty derived from TV-divergence constraints. PSPO [36] and SAPO [23] smooth the probability ratio via behavior-policy interpolation and a temperature-controlled gate, respectively. Signed LS instead smooths the weighted update ρA itself via a signed logarithmic transform.

Off-Policy Learning in Contextual Bandits. The contextual-bandit literature regularizes importance ratios rather than clipping them. Gabbianelli et al. [37] obtain tighter guarantees for importance-weighted offline learning under controlled estimation error. Aouali et al. [38] replace clipping with exponential smoothing. Sakhi et al. [15] introduce logarithmic smoothing as a concave differentiable surrogate for the importance-weighted estimator. Ryu et al. [21] recover it as a special case of a score-function family. We extend this analysis to the signed setting and prove a two-sided surrogate fidelity bound for Signed LS applied to the importance-weighted reward ρR.

7 Conclusion

Signed LS replaced ratio clipping with a signed logarithmic map of ρA that smoothly attenuated large updates rather than creating sign-dependent flat regions in the ratio surrogate. Our theory identified an explicit ratio-tail bias in hard clipping. For Signed LS, we established surrogate fidelity and regret transfer for the original importance-weighted objective. Our GRPO-style Signed LS method KRAFT outperformed clipping-based baselines on contextual-bandit benchmarks. In RLVR experiments across three LLMs, KRAFT improved GRPO's average accuracy and achieved comparable or lower reference KL and higher entropy.

As future work, we plan to extend our current analysis to multi-epoch batch reuse, off-policy actor-critic methods, and preference learning. We also aim to develop a more comprehensive understanding of KRAFT's training dynamics by theoretically analyzing why it tends to produce lower reference KL and higher policy entropy than GRPO.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) in 2026 (Grant No. RS-2026-25612350, Project Title: Development of an End-to-End Design and Validation Framework for Mathematically Principled Next-Generation AI Architectures and Learning Algorithms, contribution 45%). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH), contribution 5%; RS-2026-25616313, AI Star Fellowship(POSTECH, Interdisciplinary Program), contribution 50%). This work was supported in part by the National Science Foundation under grant CCF-2327013, Meta Platforms, Inc., and the NVIDIA Academic Grant Program using NVIDIA DGX™ Spark.

References

[1] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[2] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

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

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

[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11):3207–3260, 2013.

[6] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011.

[7] Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1): 1731–1755, 2015.

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

[9] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=2a36EMSSTp`.

[11] 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. In *Second Conference on Language Modeling*, 2025. URL `https://openreview.net/forum?id=5PAF7PAY2Y`.

[12] Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization. *arXiv preprint arXiv:2601.05242*, 2026.

[13] Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025.

[14] Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

[15] Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. *Advances in Neural Information Processing Systems*, 37:80706–80755, 2024.

[16] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

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

[18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[19] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International conference on machine learning*, pages 2139–2148. PMLR, 2016.

[20] Josiah P Hanna, Scott Niekum, and Peter Stone. Importance sampling in reinforcement learning with an estimated behavior policy. *Machine Learning*, 110(6):1267–1317, 2021.

[21] J Jon Ryu, Jeongyeol Kwon, Benjamin Koppe, and Kwang-Sung Jun. Improved offline contextual bandits with second-order bounds: Betting and freezing. In *The Thirty Eighth Annual Conference on Learning Theory*, pages 5015–5053. PMLR, 2025.

[22] Max Qiushi Lin, Jincheng Mei, Matin Aghaei, Michael Lu, Bo Dai, Alekh Agarwal, Dale Schuurmans, Csaba Szepesvari, and Sharan Vaswani. Rethinking the global convergence of softmax policy gradient with linear function approximation. *arXiv preprint arXiv:2505.03155*, 2025.

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

[24] Jacob Adkins, Michael Bowling, and Adam White. A method for evaluating hyperparameter sensitivity in reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 124820–124842, 2024.

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

[26] Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. Numinamath 7b cot. `https://huggingface.co/AI-MO/NuminaMath-7B-CoT`, 2024.

[27] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[28] Marah I Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio CT Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. Technical Report MSR-TR-2024-57, Microsoft, December 2024. URL `https://www.microsoft.com/en-us/research/publication/phi-4-technical-report/`.

[29] 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.

[30] 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.

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

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

[33] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[34] 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.

[35] Ömer Veysel Çağatan, Barış Akgün, Gözde Gül Şahin, and Xuandong Zhao. Clipping-free policy optimization for large language models. *arXiv preprint arXiv:2601.22801*, 2026.

[36] Madeleine Dwyer, Adam Sobey, and Adriane Chapman. It's not you, it's clipping: A soft trust-region via probability smoothing for llm rl. *arXiv preprint arXiv:2509.21282*, 2025.

[37] Germano Gabbianelli, Gergely Neu, and Matteo Papini. Importance-weighted offline learning done right. In *International Conference on Algorithmic Learning Theory*, pages 614–634. PMLR, 2024.

[38] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In *International Conference on Machine Learning*, pages 984–1017. PMLR, 2023.

[39] David A Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1): 100–118, 1975.

[40] Joel Tropp. Freedman's inequality for matrix martingales. 2011.

[41] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[42] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[43] Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambhuda Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft, August 2024. URL `https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/`.

[44] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[45] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[46] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.

[47] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[48] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[49] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

A Proofs and Technical Details

We prove the results from Section 4. Throughout this appendix, ρ denotes the importance ratio and β is the smoothing parameter. We use $\psi_\beta(z) = \operatorname{sgn}(z)\log(1+\beta|z|)/\beta$ and $g(y) = y^2/(1+\beta|y|)$.

A.1 Setup and Auxiliary Results

We work on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t=0}^T)$ with a fixed horizon $T \in \mathbb{N}$ and $\mathcal{F}_T \subseteq \mathcal{F}$. The history $\mathcal{F}_t$ contains the observations through round t. We use τ as the summation index when accumulating observations through round t. Contexts lie in a measurable space $(\mathcal{X}, \mathcal{B})$. The action space $\mathcal{A}$ is countable and has the discrete σ-algebra. A policy assigns an action distribution $\pi(\cdot \mid x)$ to each context x. For each action a, the map $x \mapsto \pi(a \mid x)$ is measurable.

We use the following sampling model at each round $t \in \{1, \dots, T\}$.

(i) The behavior policy μ_t and target policy π_t are predictable. For every $a \in \mathcal{A}$, their action probabilities $(\omega, x) \mapsto \mu_t(\omega, a \mid x)$ and $(\omega, x) \mapsto \pi_t(\omega, a \mid x)$ are $\mathcal{F}_{t-1} \otimes \mathcal{B}$-measurable.

(ii) The observed context $x_t : \Omega \to \mathcal{X}$ is measurable from $(\Omega, \mathcal{F}_t)$ to $(\mathcal{X}, \mathcal{B})$.

(iii) The action $a_t \in \mathcal{A}$ is $\mathcal{F}_t$-measurable and satisfies $\mathbb{P}[a_t = a \mid \mathcal{F}_{t-1} \vee \sigma(x_t)] = \mu_t(a \mid x_t)$ a.s. for every $a \in \mathcal{A}$.

(iv) The observed reward $R_t \in [-1, 1]$ is $\mathcal{F}_t$-measurable.

We write $\mathbb{E}_{t-1}[X] = \mathbb{E}[X \mid \mathcal{F}_{t-1}]$. For $X \geq 0$, this expectation may be infinite. For signed X, we take the difference of the conditional expectations of X^+ and X^- whenever the conditional expectation of $|X|$ is finite almost surely.

The importance ratio and importance-weighted feedback are

$$\rho_t := \begin{cases} \frac{\pi_t(a_t|x_t)}{\mu_t(a_t|x_t)}, & \mu_t(a_t \mid x_t) > 0, \\ 0, & \mu_t(a_t \mid x_t) = 0, \end{cases} \qquad Y_t := \rho_t R_t.$$

Because a_t is drawn from $\mu_t(\cdot \mid x_t)$, $\mu_t(a_t \mid x_t) > 0$ almost surely. Thus ρ_t and Y_t are finite almost surely and are $\mathcal{F}_t$-measurable. Under support coverage, $\mathbb{E}[\rho_t \mid \mathcal{F}_{t-1} \vee \sigma(x_t)] = 1$ almost surely.

Bounding Q_T. We first show that $Q_T = \sum_{\tau=1}^T \mathbb{E}_{\tau-1}[g(Y_\tau)]$ is finite, as required for the exponential martingales used below. Since $g(y) \leq |y|/\beta$, each term is bounded by $\mathbb{E}_{\tau-1}[|Y_\tau|]/\beta$. Under Assumptions 1 and 2, we have

$$\mathbb{E}_{\tau-1}[g(Y_\tau)] \leq \frac{1}{\beta}\mathbb{E}_{\tau-1}[|Y_\tau|] \leq \frac{1}{\beta}\mathbb{E}_{\tau-1}[\rho_\tau] = \frac{1}{\beta},$$

by the conditional mean-one property of ρ_τ. Summing over rounds gives $Q_T \leq T/\beta$ almost surely.

Identities. We exponentiate Signed LS so that its logarithm disappears, leaving a linear term in the weighted reward and a correction involving g. After taking conditional expectations, these terms give the conditional mean reward and the penalty in the deviation bound. Using both signs of the exponent lets us control overestimation and underestimation.

Lemma A.1 (Exponential identities)**.** *For every $y \in \mathbb{R}$,*

$$e^{\beta\psi_\beta(y)} = 1 + \beta y + \mathbf{1}\{y < 0\}\beta^2 g(y),$$
$$e^{-\beta\psi_\beta(y)} = 1 - \beta y + \mathbf{1}\{y \geq 0\}\beta^2 g(y),$$

and $|y| \leq \frac{1}{\beta} + 2\beta g(y)$.

Proof. For $y \geq 0$, we have $e^{\beta\psi_\beta(y)} = 1 + \beta y$. Its reciprocal is $(1+\beta y)^{-1} = 1 - \beta y + \beta^2 g(y)$. For $y < 0$, we have $e^{-\beta\psi_\beta(y)} = 1 - \beta y$. Its reciprocal is $(1-\beta y)^{-1} = 1 + \beta y + \beta^2 g(y)$. The bound on $|y|$ follows from $\beta^{-1} + 2\beta g(y) - |y| = (1+\beta^2 y^2)/[\beta(1+\beta|y|)] \geq 0$. □

We turn these exponential identities into martingale factors by dividing each exponential by its conditional mean. To justify this step, we first show that the means are finite and positive under the assumption $Q_T < \infty$.

Proposition A.1 (Conditional integrability)**.** *Under Assumption 1, suppose $Q_T < \infty$ almost surely. Then Y_τ and $\psi_\beta(Y_\tau)$ are conditionally integrable given $\mathcal{F}_{\tau-1}$ at every round. The two conditional exponential moments are finite and strictly positive almost surely, with*

$$\mathbb{E}_{\tau-1}\left[e^{\beta\psi_\beta(Y_\tau)}\right] = 1 + \beta v_\tau + \beta^2 \mathbb{E}_{\tau-1}\left[\mathbf{1}\{Y_\tau < 0\} g(Y_\tau)\right],$$
$$\mathbb{E}_{\tau-1}\left[e^{-\beta\psi_\beta(Y_\tau)}\right] = 1 - \beta v_\tau + \beta^2 \mathbb{E}_{\tau-1}\left[\mathbf{1}\{Y_\tau \geq 0\} g(Y_\tau)\right].$$

Proof. Lemma A.1 gives $\mathbb{E}_{\tau-1}[|Y_\tau|] \leq \beta^{-1} + 2\beta\mathbb{E}_{\tau-1}[g(Y_\tau)] < \infty$. Since $|\psi_\beta(y)| \leq |y|$, the same integrability holds for $\psi_\beta(Y_\tau)$. Taking conditional expectations in the two exponential identities gives the formulas for $\mathbb{E}_{\tau-1}[e^{\pm\beta\psi_\beta(Y_\tau)}]$. Each expression is finite, and each expectation is strictly positive because its integrand is strictly positive. □

Policy value. Theorem 1 compares S_t with the sum of the conditional means $v_\tau = \mathbb{E}_{\tau-1}[Y_\tau]$. When the conditional mean reward $r(x, a)$ is the same under both policies, v_τ equals the target policy's expected reward given past rounds. To prove this, we first condition on the context and action to remove the reward noise, then average over the behavior policy's actions so that the importance ratio converts this average into the target policy's value.

Proposition A.2 (Policy-value interpretation)**.** *We assume support coverage (Assumption 1) and $Q_T < \infty$ almost surely. For a fixed round τ, we define*

$$\mathcal{G}_\tau := \mathcal{F}_{\tau-1} \vee \sigma(x_\tau), \mathcal{H}_\tau := \mathcal{F}_{\tau-1} \vee \sigma(x_\tau, a_\tau).$$

Suppose the mean reward r is policy-independent and $(\mathcal{B} \otimes 2^{\mathcal{A}})$-measurable. The reward noise ξ_τ is $\mathcal{F}_\tau$-measurable and satisfies

$$R_\tau = r(x_\tau, a_\tau) + \xi_\tau \quad \textit{and} \quad \mathbb{E}[\xi_\tau \mid \mathcal{H}_\tau] = 0 \quad \textit{a.s.}$$

We also require

$$\mathbb{E}\left[|R_\tau| + |\xi_\tau| \mid \mathcal{H}_\tau\right] < \infty \textit{ a.s.}$$

For the target policy, we assume

$$\sum_{a \in \mathcal{A}} \pi_\tau(a \mid x_\tau)|r(x_\tau, a)| < \infty \textit{ a.s.}$$

and require the sum to have finite conditional expectation given $\mathcal{F}_{\tau-1}$. Then

$$v_\tau = \mathbb{E}_{\tau-1}\left[\sum_{a \in \mathcal{A}} \pi_\tau(a \mid x_\tau) r(x_\tau, a)\right].$$

Proof. Under Assumptions 1 and 2,

$$\mathbb{E}\left[|\rho_\tau R_\tau| \mid \mathcal{G}_\tau\right] \leq \mathbb{E}\left[\rho_\tau \mid \mathcal{G}_\tau\right] = 1 \quad \text{a.s.}$$

On the behavior support, ρ_τ is finite and $\mathcal{H}_\tau$-measurable. We can therefore multiply the conditional integrability bound for $|R_\tau| + |\xi_\tau|$ by ρ_τ to obtain

$$\mathbb{E}\left[|\rho_\tau R_\tau| + |\rho_\tau \xi_\tau| \mid \mathcal{H}_\tau\right] < \infty \quad \text{a.s.}$$

Averaging over the sampled action gives

$$\mathbb{E}\left[\rho_\tau |r(x_\tau, a_\tau)| \mid \mathcal{G}_\tau\right] = \sum_{a \in \mathcal{A}} \pi_\tau(a \mid x_\tau)|r(x_\tau, a)| < \infty \quad \text{a.s.}$$

It follows that $\rho_\tau \xi_\tau = \rho_\tau R_\tau - \rho_\tau r(x_\tau, a_\tau)$ is also conditionally integrable given $\mathcal{G}_\tau$.

Given $\mathcal{H}_\tau$, both ρ_τ and $r(x_\tau, a_\tau)$ are known and the noise has mean zero. Thus

$$\mathbb{E}[\rho_\tau R_\tau \mid \mathcal{H}_\tau] = \rho_\tau r(x_\tau, a_\tau) + \rho_\tau \mathbb{E}[\xi_\tau \mid \mathcal{H}_\tau] = \rho_\tau r(x_\tau, a_\tau).$$

We now average over a_τ conditional on $\mathcal{G}_\tau$. Its distribution is $\mu_\tau(\cdot \mid x_\tau)$, so

$$\begin{aligned}\mathbb{E}[Y_\tau \mid \mathcal{G}_\tau] &= \sum_{a:\mu_\tau(a|x_\tau)>0} \frac{\pi_\tau(a \mid x_\tau)}{\mu_\tau(a \mid x_\tau)} r(x_\tau, a)\mu_\tau(a \mid x_\tau) \\ &= \sum_{a \in \mathcal{A}} \pi_\tau(a \mid x_\tau) r(x_\tau, a),\end{aligned}$$

where support coverage gives $\pi_\tau(a \mid x_\tau) = 0$ for every action outside the behavior support. The sum converges absolutely because $\sum_a \pi_\tau(a \mid x_\tau)|r(x_\tau, a)| < \infty$. Conditioning once more on $\mathcal{F}_{\tau-1}$ gives the formula for v_τ. □

A.2 Proofs of Main Results

We construct two exponential martingales, one for each direction of the estimation error. Ville's inequality bounds each martingale over all rounds at once, and the identities above turn these bounds into inequalities for $S_t - V_t$ and $V_t - S_t$.

Under Assumptions 1 and 2, the bound in Appendix A.1 gives $Q_T < \infty$ a.s. To apply Ville's inequality, we define

$$Z_\tau^+ := \frac{e^{\beta\psi_\beta(Y_\tau)}}{\mathbb{E}_{\tau-1}[e^{\beta\psi_\beta(Y_\tau)}]} \quad \text{and} \quad Z_\tau^- := \frac{e^{-\beta\psi_\beta(Y_\tau)}}{\mathbb{E}_{\tau-1}[e^{-\beta\psi_\beta(Y_\tau)}]},$$

and set $M_t^\pm = \prod_{\tau=1}^t Z_\tau^\pm$ with $M_0^\pm = 1$. Proposition A.1 ensures that both denominators are finite and positive. Each factor $Z_\tau^\pm$ is nonnegative and $\mathcal{F}_\tau$-measurable with conditional mean one given $\mathcal{F}_{\tau-1}$. Thus each product is a nonnegative martingale with mean one.

Upper bound on $S_t - V_t$. By Ville's inequality, $\mathcal{E}_\delta^+ = \{\sup_{0\le t\le T} M_t^+ \le 1/\delta\}$ has probability at least $1-\delta$. On this event, for every $t \le T$ we have

$$\sum_{\tau=1}^t \beta\psi_\beta(Y_\tau) - \sum_{\tau=1}^t \log \mathbb{E}_{\tau-1}\left[e^{\beta\psi_\beta(Y_\tau)}\right] = \log M_t^+ \le \log\left(\frac{1}{\delta}\right).$$

For each τ, we write $u = \beta v_\tau + \beta^2 \mathbb{E}_{\tau-1}[\mathbf{1}\{Y_\tau < 0\} g(Y_\tau)]$. Proposition A.1 gives $1 + u = \mathbb{E}_{\tau-1}[e^{\beta\psi_\beta(Y_\tau)}] > 0$. The inequality $\log(1+u) \le u$ therefore yields

$$\log \mathbb{E}_{\tau-1}\left[e^{\beta\psi_\beta(Y_\tau)}\right] \le \beta v_\tau + \beta^2 \mathbb{E}_{\tau-1}\left[\mathbf{1}\{Y_\tau < 0\} g(Y_\tau)\right].$$

On $\mathcal{E}_\delta^+$, substituting this estimate into $\log M_t^+$ and dividing by β gives, for every $1 \le t \le T$,

$$S_t - V_t \le \beta Q_t^- + \frac{1}{\beta}\log\left(\frac{1}{\delta}\right).$$

Upper bound on $V_t - S_t$. To bound $V_t - S_t$, we use M_t^- and the identity for $\mathbb{E}_{\tau-1}[e^{-\beta\psi_\beta(Y_\tau)}]$. On $\mathcal{E}_\delta^- = \{\sup_{0\le t\le T} M_t^- \le 1/\delta\}$, which also has probability at least $1-\delta$, we have for every $1 \le t \le T$

$$V_t - S_t \le \beta Q_t^+ + \frac{1}{\beta}\log\left(\frac{1}{\delta}\right).$$

Two-sided bound. With failure probability $\delta/2$ in each direction, both bounds hold on an event of probability at least $1-\delta$. On this event,

$$|S_t - V_t| \le \beta(Q_t^- + Q_t^+) + \frac{1}{\beta}\log\left(\frac{2}{\delta}\right) = \beta Q_t + \frac{1}{\beta}\log\left(\frac{2}{\delta}\right) \text{ for every } t \in \{1, \dots, T\}.$$

This proves Theorem 1. □

Proposition A.3 (Importance ratios with a finite p-th moment)**.** *Under Assumptions 1 and 2, suppose $\mathbb{E}_{\tau-1}[\rho_\tau^p] \le M_p$ almost surely at every round for some $1 < p \le 2$ and finite M_p. For each fixed $\beta > 0$ and $0 < \delta < 1$, the following bound holds simultaneously for all $1 \le t \le T$ with probability at least $1-\delta$.*

$$|S_t - V_t| \le M_p t\, \beta^{p-1} + \frac{\log(2/\delta)}{\beta}.$$

If we choose $\beta_T = \left(\frac{\log(2/\delta)}{(p-1)M_p T}\right)^{1/p}$ before sampling, then at $t = T$ we have

$$\frac{|S_T - V_T|}{T} \le \frac{p}{(p-1)^{(p-1)/p}} M_p^{1/p} \left(\frac{\log(2/\delta)}{T}\right)^{(p-1)/p}$$

with probability at least $1-\delta$.

Proof. For $u > 0$, we have $u^{2-p} \le 1 + u$. Taking $u = \beta|y|$ gives $g(y) \le \beta^{p-2}|y|^p$, which also holds at $y = 0$. Since $|Y_\tau| \le \rho_\tau$, we have $Q_t \le t\beta^{p-2}M_p$. Substituting into Theorem 1 proves the uniform bound. At $t = T$, its right-hand side is minimized when $(p-1)M_pT\beta^p = \log(2/\delta)$. For $p < 2$, the moment assumption allows importance ratios with infinite variance. □

From bandit actions to complete responses. We can apply Theorem 1 to the reward of a complete response by treating the response as one bandit action. Under either policy, the probability of a response is the product of its next-token probabilities. Dividing the target probability of the response by its behavior probability gives the product of the token ratios.

The same calculation applies to a trajectory of length H. In round t, both policies are chosen using only past rounds, and the behavior policy μ_t generates the trajectory. The initial context has the same distribution under both policies. Given the history and action, the environment draws the next observation and reward from the same joint distribution under either policy. If $h_{t,j}$ is the history before action $a_{t,j}$, the trajectory has importance weight

$$W_t = \prod_{j=1}^{H} \frac{\pi_t(a_{t,j} \mid h_{t,j})}{\mu_t(a_{t,j} \mid h_{t,j})}.$$

Let G_t be the return and J_t its conditional mean under π_t given past rounds.

Corollary A.1 (Fidelity for complete trajectories)**.** *Suppose $|G_t| \le 1$. Conditional on past rounds, we assume that the trajectory distribution under the target policy is absolutely continuous with respect to that under the behavior policy. For any fixed $\beta > 0$ and $0 < \delta < 1$, the following inequality holds simultaneously for all $1 \le t \le T$ with probability at least $1 - \delta$.*

$$\left| \sum_{\tau=1}^{t} \big(\psi_\beta(W_\tau G_\tau) - J_\tau\big) \right| \le \beta \sum_{\tau=1}^{t} \mathbb{E}_{\tau-1}[g(W_\tau G_\tau)] + \frac{\log(2/\delta)}{\beta}.$$

Proof. The environment terms cancel in the likelihood ratio between the target and behavior trajectory distributions, leaving the product W_τ of the action probability ratios. Hence $\mathbb{E}_{\tau-1}[W_\tau G_\tau] = J_\tau$ and $\mathbb{E}_{\tau-1}[W_\tau] = 1$. The proof of Theorem 1 applies to $Y_\tau = W_\tau G_\tau$ since $\mathbb{E}_{\tau-1}[|Y_\tau|] \le 1$. □

Per-token advantages in GRPO. In KRAFT, each token has its own importance ratio. For a single response, let $\mathcal{F}_\ell$ be the history generated by the prompt and the first ℓ tokens. For a bounded $\mathcal{F}_\ell$-measurable weight A_ℓ, the proof of Theorem 1 applies to $Y_\ell = \rho_\ell A_\ell$ and compares $\sum_{j=1}^{\ell} \psi_\beta(Y_j)$ with $\sum_{j=1}^{\ell} \mathbb{E}_{j-1}[Y_j]$. If $|A_\ell| \le A_{\max}$, then $g(y) \le |y|/\beta$ and $\mathbb{E}_{\ell-1}[\rho_\ell] = 1$ imply $Q_\ell \le \ell A_{\max}/\beta$. GRPO uses the same advantage A_i for every token in response i. Computing A_i requires the rewards of all responses in the group. If rewards lie in $[-1, 1]$ and the normalization denominator is at least $\varepsilon_0 > 0$, then $|A_i| \le 2/\varepsilon_0$ because $|R_i - \bar{R}| \le 2$. Even with this bound, the prompt and the first ℓ tokens of response i generally do not determine A_i. Thus the measurability condition in the token-level argument need not hold. If training reuses rollouts, the updated policy also depends on the rollout being evaluated. At a fixed policy, subtracting the same KL penalty from the smoothed and importance-weighted objectives preserves their difference.

Regret transfer. We write the importance-weighted regret as the surrogate regret plus two approximation errors, one for the comparator and one for the learner. These errors enter with opposite signs, so the regret bound uses a different direction of the fidelity bound for each policy.

For the fixed comparator $\widetilde{\pi}$, the upper bound on $V_T(\widetilde{\pi}) - S_T(\widetilde{\pi})$ gives, with probability at least $1 - \delta/2$,

$$V_T(\widetilde{\pi}) - S_T(\widetilde{\pi}) \le \beta Q_T^+(\widetilde{\pi}) + \frac{1}{\beta} \log\left(\frac{2}{\delta}\right).$$

For the learner, we need an upper bound on $S_T - V_T$. With probability at least $1 - \delta/2$,

$$S_T - V_T \le \beta Q_T^- + \frac{1}{\beta} \log\left(\frac{2}{\delta}\right).$$

Both bounds hold on an event of probability at least $1-\delta$. Adding them to the surrogate regret gives

$$\begin{aligned} R_T^{\mathsf{IW}}(\widetilde{\pi}) = V_T(\widetilde{\pi}) - V_T &= (V_T(\widetilde{\pi}) - S_T(\widetilde{\pi})) + (S_T(\widetilde{\pi}) - S_T) + (S_T - V_T) \\ &\le R_T^{\mathsf{surr}}(\widetilde{\pi}) + \beta(Q_T^+(\widetilde{\pi}) + Q_T^-) + \frac{2}{\beta}\log\left(\frac{2}{\delta}\right). \end{aligned}$$

□

A.3 Proofs for Clipping

Lemma A.2 (Clipping bias bound)**.** *Under Assumption 1, for every round τ,*

$$|b_\tau(\kappa)| \le \mathbb{E}_{\tau-1}\left[(\rho_\tau - \kappa)_+ |R_\tau|\right] \le \mathbb{E}_{\tau-1}\left[\frac{2\rho_\tau}{\kappa+\rho_\tau}(\rho_\tau - \kappa)_+ |R_\tau|\right].$$

Proof. The triangle inequality gives the first bound. When $\rho_\tau \ge \kappa$, the factor $2\rho_\tau/(\kappa+\rho_\tau)$ is at least one. For $\rho_\tau < \kappa$, the positive part vanishes. This proves the second bound. □

Proof of Theorems 2 and 3. For both clipping rules, we separate the estimation error into a centered sum and the bias introduced by clipping. The clipped rewards are bounded, so we control the centered sum with an exponential supermartingale and then add the bias bound.

We write $Z_\tau = \min\{\rho_\tau, \kappa\}R_\tau$ for hard clipping and $Z_\tau = \rho_\tau R_\tau/(1+\rho_\tau/\kappa)$ for soft clipping. We define the centered increments $X_\tau = Z_\tau - \mathbb{E}_{\tau-1}[Z_\tau]$ and their cumulative conditional variance $W_t = \sum_{\tau\le t}\mathbb{E}_{\tau-1}[X_\tau^2]$. In both cases, $|X_\tau| \le 2\kappa$ and $\mathbb{E}_{\tau-1}[X_\tau] = 0$. For fixed $0 < \lambda < 3/(2\kappa)$, expanding the exponential and using $j! \ge 2\cdot 3^{j-2}$ for $j \ge 2$ gives

$$\begin{aligned} \mathbb{E}_{\tau-1}[e^{\lambda X_\tau}] &\le 1 + \frac{\lambda^2\mathbb{E}_{\tau-1}[X_\tau^2]}{2}\sum_{j=0}^{\infty}(2\kappa\lambda/3)^j \\ &\le \exp\left(\frac{\lambda^2\mathbb{E}_{\tau-1}[X_\tau^2]}{2(1-2\kappa\lambda/3)}\right). \end{aligned}$$

The process $M_t = \exp\{\lambda\sum_{\tau\le t}X_\tau - \lambda^2 W_t/[2(1-2\kappa\lambda/3)]\}$ is a nonnegative supermartingale starting at one. We apply Ville's inequality to the processes for X_τ and $-X_\tau$, with failure probability $\delta/2$ for each. With probability at least $1-\delta$, for every $1 \le t \le T$ we have

$$\left|\sum_{\tau\le t}X_\tau\right| \le \frac{\lambda W_t}{2(1-2\kappa\lambda/3)} + \frac{1}{\lambda}\log\frac{2}{\delta}.$$

For either clipping rule, $\sum_{\tau\le t}Z_\tau - V_t = \sum_{\tau\le t}X_\tau - B_t$. Adding the absolute bias $|B_t|$ to the bound on the centered sum proves the result. □

Choice of λ. Freedman's inequality uses a fixed upper bound on the variance [39, 40]. To choose λ using the observed variance, we need a bound that holds for all candidate values of λ at once. Mixture and stitching methods give such bounds [41].

Proof of the comparison and bias limit. For a fixed sample size, we compare the bounds by applying the Signed LS fidelity result and separating the soft-clipping error into sampling error and bias. Bernstein's inequality controls the sampling error, while the explicit bias formula shows both how the clipping level affects the bound and what happens as that level tends to infinity.

Proposition A.4 (Fixed-sample comparison with soft clipping)**.** *We draw n i.i.d. samples from a fixed behavior policy and evaluate a fixed target policy satisfying support coverage. For rewards satisfying $0 \le R \le 1$, let $m_2 = \mathbb{E}[(\rho R)^2]$ and assume $h_2 = \mathbb{E}[\rho^2 R] < \infty$. Then, for any fixed $\beta, \kappa > 0$ and $0 < \delta < 1$, the following two inequalities hold jointly with probability at least $1-\delta$.*

$$\left|\frac{S_n}{n} - \mathbb{E}[\rho R]\right| \le \beta m_2 + \frac{\log(4/\delta)}{n\beta},$$

$$\left|\frac{S_n^{\mathsf{sclip}}(\kappa)}{n} - \mathbb{E}[\rho R]\right| \le \frac{h_2}{\kappa} + \sqrt{\frac{2m_2\log(4/\delta)}{n}} + \frac{2\kappa\log(4/\delta)}{3n}.$$

For i.i.d. samples, $V_n = n\mathbb{E}[\rho R]$ and $Q_n \le nm_2$ because $g(y) \le y^2$. Applying Theorem 1 with failure probability $\delta/2$ gives the Signed LS bound. For soft clipping, the reward $Z = \rho R/(1+\rho/\kappa)$ satisfies

$$0 \le \mathbb{E}[\rho R] - \mathbb{E}[Z] = \mathbb{E}\left[\frac{\rho^2 R}{\kappa+\rho}\right] \le \frac{h_2}{\kappa}, \qquad \mathrm{Var}(Z) \le \mathbb{E}[Z^2] \le m_2.$$

The bound $0 \le Z \le \kappa$ implies $|Z - \mathbb{E}[Z]| \le \kappa$, and Bernstein's inequality gives

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbb{E}[Z]\right| \le \sqrt{\frac{2m_2\log(4/\delta)}{n}} + \frac{2\kappa\log(4/\delta)}{3n},$$

with failure probability at most $\delta/2$. Adding the bias bound h_2/κ gives the soft-clipping inequality. By a union bound, this inequality and the Signed LS inequality hold together with probability at least $1-\delta$. When $m_2, h_2 > 0$, minimizing over β gives $2\sqrt{m_2\log(4/\delta)/n}$ for Signed LS. Minimizing over κ gives $\sqrt{2m_2\log(4/\delta)/n} + 2\sqrt{2h_2\log(4/\delta)/(3n)}$ for soft clipping. Since $R^2 \le R$, we have $m_2 \le h_2$.

As $\kappa \to \infty$, the bias term $\rho_\tau^2 R_\tau/(\kappa+\rho_\tau)$ tends to zero. Its absolute value is bounded by $\rho_\tau|R_\tau|$, whose conditional mean is at most one. Conditional dominated convergence gives $B_t^{\mathsf{sclip}}(\kappa) \to 0$ for each fixed t. If rewards are nonnegative, $\kappa\rho_\tau^2 R_\tau/(\kappa+\rho_\tau)$ increases to $\rho_\tau^2 R_\tau$. By conditional monotone convergence, $\kappa B_t^{\mathsf{sclip}}(\kappa)$ tends to $\sum_{\tau\le t}\mathbb{E}_{\tau-1}[\rho_\tau^2 R_\tau]$. This limit may be infinite.

A.4 Proofs for the SGD Analysis

For a fixed prompt x, we draw $G \ge 2$ independent action–reward pairs (a_i, R_i) under π_θ. We compute advantages from the rewards of the sampled group as follows:

$$A_i = \frac{R_i - \bar{R}}{\sqrt{G^{-1}\sum_j (R_j - \bar{R})^2 + \varepsilon_0}},$$

where $\bar{R} = G^{-1}\sum_i R_i$ and $\varepsilon_0 \ge 0$. We set all advantages to zero when the group rewards are equal. To compute the two gradients, we hold the advantages and the behavior policy fixed and evaluate at unit importance ratio. This gives

$$\widehat{g}^{\mathsf{GRPO}}(\theta) = \frac{1}{G}\sum_i A_i \nabla_\theta \log \pi_\theta(a_i \mid x),$$

$$\widehat{g}^{\mathsf{KRAFT}}(\theta) = \frac{1}{G}\sum_i \frac{A_i}{1+\beta|A_i|}\nabla_\theta \log \pi_\theta(a_i \mid x).$$

At step k, each method samples its own group from its current policy. With learning rates η for GRPO and $\widetilde{\eta}$ for KRAFT, the updates are

$$\theta_{k+1}^{\mathsf{GRPO}} = \theta_k^{\mathsf{GRPO}} + \eta\widehat{g}_k^{\mathsf{GRPO}}(\theta_k^{\mathsf{GRPO}}),$$
$$\theta_{k+1}^{\mathsf{KRAFT}} = \theta_k^{\mathsf{KRAFT}} + \widetilde{\eta}\widehat{g}_k^{\mathsf{KRAFT}}(\theta_k^{\mathsf{KRAFT}}).$$

Proof. We first condition on the number of successful responses to show that the two expected updates point in the same direction, with KRAFT taking a smaller step. We then expand the expected reward after N updates around zero learning rate and invert this relation to compare the learning rates needed for the same small reward gain.

Let $p = J_x(\theta) \in (0,1)$ be the success probability and $s_\theta(a) = \nabla_\theta \log \pi_\theta(a \mid x)$ the policy score. Differentiating the sum of the action probabilities gives $\mathbb{E}_\theta[s_\theta(a) \mid x] = 0$. Differentiating p gives $\mathbb{E}_\theta[R s_\theta(a) \mid x] = \nabla p$ because the conditional reward distribution given (x,a) does not depend on θ. Since R is binary, the mean score for each outcome is

$$\mathbb{E}_\theta[s_\theta(a) \mid x, R=1] = \frac{\nabla p}{p}, \qquad \mathbb{E}_\theta[s_\theta(a) \mid x, R=0] = -\frac{\nabla p}{1-p}.$$

Suppose a group contains k successes with $1 \le k < G$. The normalization denominator is $d_k = \sqrt{k(G-k)/G + \varepsilon_0}$. Each success has advantage $u_k = (G-k)/(Gd_k)$ and each failure has advantage $-v_k = -k/(Gd_k)$. By independence of the action–reward pairs, conditioning on the other

rewards does not change the mean of $s_\theta(a_i)$ given x and R_i. Using $ku_k = (G-k)v_k$, we obtain

$$\mathbb{E}_\theta[\widehat{g}^{\mathsf{GRPO}} \mid x, k] = \frac{ku_k}{Gp(1-p)}\nabla p,$$
$$\mathbb{E}_\theta[\widehat{g}^{\mathsf{KRAFT}} \mid x, k] = \frac{ku_k}{Gp(1-p)}\left(\frac{1-p}{1+\beta u_k} + \frac{p}{1+\beta v_k}\right)\nabla p.$$

For each $1 \le k < G$, smoothing multiplies the conditional expected update by a factor strictly between zero and one. Groups with $k = 0$ or $k = G$ contribute zero. To average over $k \sim \mathrm{Binomial}(G, p)$, we write $w_k = \binom{G}{k}p^k(1-p)^{G-k}ku_k/[Gp(1-p)]$. Let b denote the coefficient of $\nabla J_x(\theta)$ in the expected GRPO update, and let c be the factor by which KRAFT scales that update. Averaging over the number of successes gives

$$b = \sum_{k=1}^{G-1} w_k > 0, \qquad c = \frac{1}{b}\sum_{k=1}^{G-1} w_k\left(\frac{1-p}{1+\beta u_k} + \frac{p}{1+\beta v_k}\right) \in (0,1).$$

Thus $\mathbb{E}_\theta[\widehat{g}^{\mathsf{GRPO}} \mid x] = b\nabla J_x(\theta)$ and $\mathbb{E}_\theta[\widehat{g}^{\mathsf{KRAFT}} \mid x] = cb\nabla J_x(\theta)$.

For either method, let $H(\theta)$ be its stochastic gradient and $F_n(\theta, \eta)$ the expectation of $J_x(\theta_n)$ over the groups sampled in n updates, starting from θ with learning rate η. We first differentiate F_N at $\eta = 0$. On a small compact neighborhood of θ_0, the action probabilities are positive and the first two derivatives of H for each possible group are bounded. Since N is fixed and only finitely many groups are possible, all paths starting sufficiently close to θ_0 stay in that neighborhood for small enough $|\eta|$. Summing over these paths shows that F_n is C^2 for $n \le N$. At zero learning rate, $F_n(\theta, 0) = J_x(\theta)$. Conditioning on the first group and differentiating gives

$$F_{n+1}(\theta, \eta) = \mathbb{E}_\theta[F_n(\theta + \eta H(\theta), \eta) \mid x],$$
$$\partial_\eta F_{n+1}(\theta, 0) = \partial_\eta F_n(\theta, 0) + \langle \nabla J_x(\theta), \mathbb{E}_\theta[H(\theta) \mid x]\rangle.$$

At θ_0, summing this recursion over N steps gives the derivatives $Nb\|\nabla J_x(\theta_0)\|^2$ for GRPO and $Ncb\|\nabla J_x(\theta_0)\|^2$ for KRAFT. Both are positive by the hypothesis $\nabla J_x(\theta_0) \neq 0$. The inverse function theorem therefore gives a locally unique positive learning rate for each sufficiently small gain $\Delta > 0$. The two inverse functions are C^2. Writing $D = Nb\|\nabla J_x(\theta_0)\|^2$, the inverse expansions give $\eta(\Delta) = \Delta/D + O(\Delta^2)$ and $\widetilde{\eta}(\Delta) = \Delta/(cD) + O(\Delta^2)$. Hence $\widetilde{\eta}(\Delta)/\eta(\Delta) = 1/c + O(\Delta)$ for fixed N. □

B Contextual Bandit Experiment Details

B.1 Bandit Environment and Objectives

Each round draws a context $x \in \mathbb{R}^{d_1}$ from $\mathcal{D}_X$ and selects from a finite action set $\mathcal{A} \subset \mathbb{R}^{d_2}$. The action features are drawn as $a_k \sim \mathcal{N}(0, I_{d_2})$ and held fixed. For a selected action a, the observed reward is

$$r = \tanh(\langle \phi(x,a), \mu^* \rangle + n),$$
$$\phi(x,a) := \mathrm{vec}(xa^\top) \in \mathbb{R}^{d_1 d_2},$$
$$n \sim \mathcal{N}(0, \alpha^2),$$
$$\mu^* = \left(\frac{s}{\|\widetilde{\theta}\|}\right)\widetilde{\theta},$$
$$\widetilde{\theta} \sim \mathcal{N}(0, I_d).$$

The learner is a softmax policy $\pi_\theta(a|x) \propto \exp(\langle \phi(x,a), \theta \rangle)$. With $r^*(x, a_k) := \tanh(\langle \phi(x, a_k), \mu^* \rangle)$, we evaluate expected noiseless reward on 1,000 held-out contexts:

$$\mathbb{E}_x\left[\sum_k \pi_\theta(a_k|x) r^*(x, a_k)\right].$$

At each iteration we draw M actions per context from $\pi_{\theta_{\text{old}}}$ and form group-normalized advantages

$$A_m = \frac{r_m - \bar{r}}{\text{std}(r_{1:M}) + \varepsilon_0},$$

where ε_0 is a small numerical constant. For $\rho_i := \frac{\pi_\theta(a_i|x_i)}{\pi_{\theta_{\text{old}}}(a_i|x_i)}$, the compared per-action losses are

$$\begin{aligned}
\mathcal{L}_i^{\text{base}} &= -\rho_i A_i && \text{(Baseline)} \\
\mathcal{L}_i^{\text{PPO-clip}} &= -\min\big(\rho_i A_i, \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)A_i\big) && \text{(PPO Clipping)} \\
\mathcal{L}_i^{\text{clip}} &= -\min(\rho_i, 1+\varepsilon)A_i && \text{(Clipping)} \\
\mathcal{L}_i^{\text{SAPO-sym}} &= -\sigma\big(\tau(\rho_i - 1)\big) \cdot \tfrac{4}{\tau} \cdot A_i && \text{(SAPO-sym)} \\
\mathcal{L}_i^{\text{ratio}} &= -\frac{1}{\beta}\log(1+\beta\rho_i)A_i && \text{(Ratio-only LS)} \\
\mathcal{L}_i^{\text{KRAFT}} &= -\frac{1}{\beta}\text{sgn}(A_i)\log\big(1+\beta\rho_i|A_i|\big) && \text{(KRAFT)}
\end{aligned}$$

Ratio-only LS smooths ρ_i. KRAFT applies Signed LS to the full weighted update $\rho_i A_i$.

B.2 Hyperparameter Grids

We use SGD and search a 1–2–5 grid ($c \times 10^k$, $c \in \{1, 2, 5\}$) over the ranges in Table 3. All selected hyperparameters lie strictly inside their search ranges.

Table 3: Hyperparameter search ranges for each method. λ denotes ε (PPO Clipping and Clipping), τ (SAPO-sym), or β (Ratio-only LS and KRAFT).

Method	η	λ
Baseline	$\{10^{-3}, 2\times10^{-3}, \ldots, 2\times10^{-2}, 5\times10^{-2}\}$	—
PPO Clipping	$\{10^{-2}, 2\times10^{-2}, \ldots, 10^{-1}, 2\times10^{-1}\}$	$\{5\times10^{-4}, 10^{-3}, \ldots, 5, 10\}$
Clipping	$\{10^{-2}, 2\times10^{-2}, \ldots, 10^{-1}, 2\times10^{-1}\}$	$\{10^{-5}, 2\times10^{-5}, \ldots, 5, 10\}$
SAPO-sym	$\{10^{-2}, 2\times10^{-2}, \ldots, 10^{-1}, 2\times10^{-1}\}$	$\{5, 10, \ldots, 5\times10^5, 10^6\}$
Ratio-only LS	$\{5\times10^{-2}, 10^{-1}, \ldots, 10^5, 2\times10^5\}$	$\{1, 2, \ldots, 5\times10^6, 10^7\}$
KRAFT	$\{2\times10^{-2}, 5\times10^{-2}, \ldots, 10, 20\}$	$\{1, 2, \ldots, 5\times10^2, 10^3\}$

Each configuration is evaluated over 200 independent random seeds. For each seed, we resample the environment independently, including the action features a_k, the true parameter μ^*, and the test contexts. We also re-initialize the policy and optimizer for each seed. For each seed, all methods and hyperparameter settings use the same environment and initial policy.

B.3 Default Experimental Settings

Unless otherwise stated, all experiments use the settings in Table 4. Each training epoch draws one context (batch size 1), samples M actions from $\pi_{\theta_{\text{old}}}$, and performs E gradient updates on the sampled group, so 1,000 epochs correspond to 3,000 update steps in Figure 3. The sweeps in Figure 4 and Table 6 vary one of d_1, d_2, α, $|\mathcal{A}|$, or E while holding the others fixed.

Table 4: Default experimental settings for the contextual bandit experiments.

Hyperparameter	**Symbol**	**Value**
Context dimension	d_1	10
Action feature dimension	d_2	4
Parameter dimension	$d = d_1 d_2$	40
Number of actions	$\|\mathcal{A}\|$	100
Noise level	α	5.0
Norm scale	s	1.0
Training epochs	—	1,000
Batch size	—	1
Group size	M	5
PPO epochs	E	3

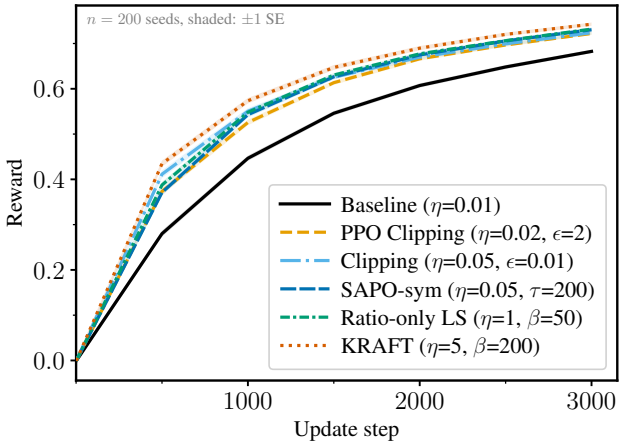


Figure 3: Test reward vs. update step (mean $\pm$ 1 SE over 200 seeds). Each method uses its best (η, λ) selected at the final step.

Table 5: Test reward for every 500 steps (mean $\pm$ 1 SE over 200 seeds). Each method uses its best (η, λ) selected at the final step. **Bold** marks the best value in each row. In every row, the gap between KRAFT and each other method is significant at the 0.05 level under a paired t test.

Step	Baseline	PPO Clipping	Clipping	SAPO-sym	Ratio-only LS	KRAFT
500	0.2800 $\pm$0.0068	0.3734 $\pm$0.0069	0.4113 $\pm$0.0084	0.3710 $\pm$0.0071	0.3875 $\pm$0.0068	**0.4347 $\pm$0.0069**
1000	0.4465 $\pm$0.0059	0.5256 $\pm$0.0062	0.5505 $\pm$0.0066	0.5428 $\pm$0.0061	0.5482 $\pm$0.0053	**0.5736 $\pm$0.0061**
1500	0.5459 $\pm$0.0051	0.6135 $\pm$0.0048	0.6255 $\pm$0.0052	0.6265 $\pm$0.0050	0.6304 $\pm$0.0043	**0.6471 $\pm$0.0054**
2000	0.6072 $\pm$0.0047	0.6663 $\pm$0.0044	0.6698 $\pm$0.0048	0.6747 $\pm$0.0044	0.6772 $\pm$0.0039	**0.6897 $\pm$0.0047**
2500	0.6481 $\pm$0.0045	0.6968 $\pm$0.0039	0.6992 $\pm$0.0047	0.7056 $\pm$0.0041	0.7063 $\pm$0.0038	**0.7201 $\pm$0.0040**
3000	0.6826 $\pm$0.0039	0.7222 $\pm$0.0034	0.7239 $\pm$0.0041	0.7303 $\pm$0.0036	0.7318 $\pm$0.0036	**0.7426 $\pm$0.0036**

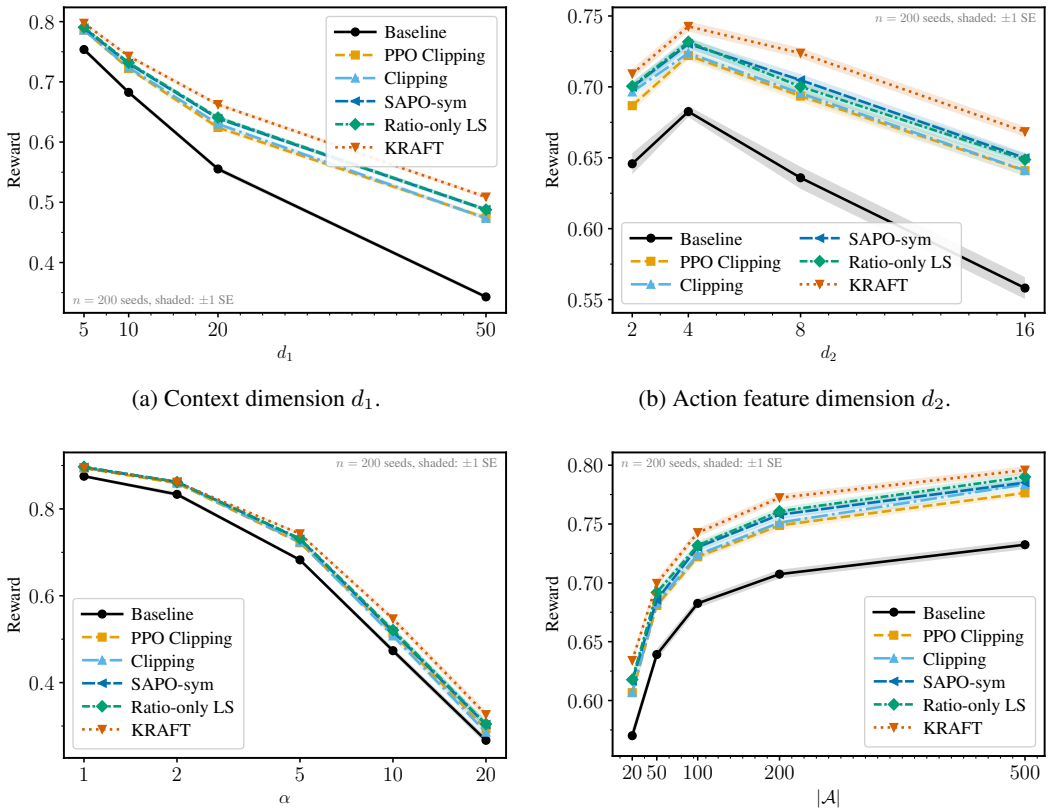
B.4 Additional Results

Learning Curves. Figure 3 and Table 5 show that all five non-baseline methods exceed the baseline throughout training. KRAFT has the largest mean reward at every step. At early steps, Clipping has the largest mean reward among the non-KRAFT baselines. Ratio-only LS overtakes Clipping at step 1500 and has the second-largest mean reward from then on. The gap between KRAFT and every other method, including Ratio-only LS, is significant at the 0.05 level under a paired t test at every step. PPO Clipping and Clipping reach comparable final rewards. SAPO-sym reaches a slightly higher final reward but remains below both LS variants.

Robustness Across Experimental Settings. Figure 4 shows final test reward across four sweeps: context dimension d_1, action-feature dimension d_2, reward noise α, and number of actions $|\mathcal{A}|$. KRAFT matches or beats all baselines across all settings. The margin over clipping-based methods is modest at small context dimension (Figure 4 (a)) and low noise (Figure 4 (c)), and widens as context dimension and noise increase.

Tuning Sensitivity Details. Each sweep plays the role of the multiple environments in the original protocol [24]. The five sweeps vary context dimension $d_1 \in \{5, 10, 20, 50\}$, action feature dimension $d_2 \in \{2, 4, 8, 16\}$, reward noise $\alpha \in \{1, 2, 5, 10, 20\}$, number of actions $|\mathcal{A}| \in \{20, 50, 100, 200, 500\}$, and PPO epochs $E \in \{2, 3, 4, 5\}$. Each setting is evaluated over 200 independent random seeds. Table 6 and Figure 1 show that KRAFT has the smallest overall value (0.0013) and leads in four of five sweeps. SAPO-sym has the lowest sensitivity on the PPO-epoch sweep (0.0018).

Reward versus Policy Entropy. We examine whether KRAFT's reward gains come at the cost of lower policy entropy by comparing it with PPO Clipping at reward noise levels $\alpha \in \{5, 10\}$.



(a) Context dimension d_1.

(b) Action feature dimension d_2.

(c) Reward noise α.

(d) Number of actions $|\mathcal{A}|$.

Figure 4: Final test reward across problem settings (mean $\pm$ 1 SE over 200 seeds). Best (η, λ) per setting: $\lambda = \varepsilon$ (PPO Clipping and Clipping), τ (SAPO-sym), β (Ratio-only LS and KRAFT).

Table 6: Exact values corresponding to Figure 1. **Bold**: best per row.

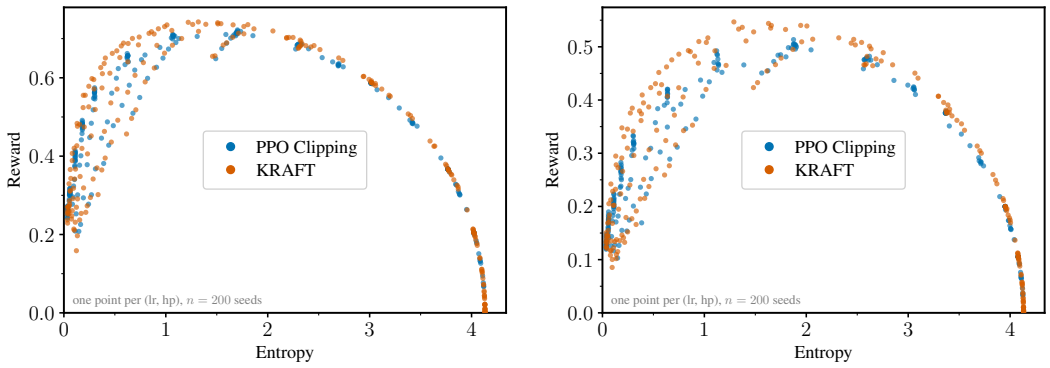
Sweep	PPO Clipping	Clipping	SAPO-sym	Ratio-only LS	KRAFT
Context dim d_1	0.0043	0.0016	0.0026	0.0053	**0.0012**
Action dim d_2	0.0030	0.0031	0.0029	0.0070	**0.0005**
Noise α	0.0047	0.0034	0.0032	0.0054	**0.0004**
Num actions $\lvert\mathcal{A}\rvert$	0.0049	0.0071	0.0022	0.0035	**0.0020**
PPO Epochs E	0.0052	0.0058	**0.0018**	0.0024	0.0020
Overall	0.0036	0.0049	0.0031	0.0047	**0.0013**

For each method, we evaluate 221 (η, λ) configurations on a 1–2–5 grid over $\eta \in [10^{-3}, 10]$ and $\lambda \in [10^{-3}, 2 \times 10^2]$, where λ denotes ε for PPO Clipping and β for KRAFT. Figure 5 plots final test reward against final policy entropy, both averaged over 200 seeds.

Both methods reach their highest mean rewards at intermediate entropy levels and perform similarly at high entropy. At comparable intermediate entropy levels, however, KRAFT generally achieves higher rewards, especially at $\alpha = 10$. This suggests that KRAFT's gains do not simply reflect lower policy entropy. In RLVR, KRAFT also achieves higher average accuracy with higher entropy (Figure 2).

C Numerical Comparison of Soft Clipping and Signed LS

The upper bounds in Section 4.1 do not determine which estimator has lower actual error. We therefore compare soft clipping and Signed LS numerically on a two-context instance.



(a) Reward noise $\alpha = 5$.

(b) Reward noise $\alpha = 10$.

Figure 5: Final test reward versus final policy entropy for PPO Clipping and KRAFT. Each point represents one of 221 (η, λ) configurations per method, with both quantities averaged over 200 seeds.

We consider contexts x_1, x_2 with $\Pr(x_2) = p$ and actions a_1, a_2. The behavior policy satisfies $\mu(a_2|x_1) = m$ and $\mu(a_2|x_2) = q$. Rewards are deterministic, with $R(x_1, a_2) = c$, $R(x_2, a_2) = 1$, and zero otherwise. The target policy always selects a_2, so its value is $v := \mathbb{E}[\rho R] = (1-p)c + p$. For each sample size n, we set $p = q = m = c = n^{-1/2}$.

Both rewarding pairs have the same importance ratio $\rho = \sqrt{n}$, but their weighted rewards are 1 and $\sqrt{n}$, respectively. They contribute comparably to the target value, yet the rare pair dominates the variance as n grows. Soft clipping attenuates both by the same factor. Signed LS instead transforms ρR directly: since $\psi_\beta(y)/y$ decreases with $y > 0$, it shrinks the rare contribution by a larger proportion.

For each n, we draw n i.i.d. samples under the behavior policy and compute the two estimators. Since rewards are nonnegative, these take the form

$$\widehat{v}_n^{\mathsf{sclip}}(\kappa) = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_i R_i}{1 + \rho_i/\kappa}, \qquad \widehat{v}_n^{\mathsf{LS}}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\beta}\log\bigl(1 + \beta\rho_i R_i\bigr).$$

We measure the scaled absolute error $\sqrt{n}\,|v - \widehat{v}|$ over 3,000 independent trials. For each n, we tune each estimator by minimizing the mean scaled error over 1,000 logarithmically spaced values in $[10^{-5}, 10]$ for $1/\kappa$ and β, respectively. All selected values lie strictly inside the search range. This oracle comparison uses the same trials for tuning and evaluation. Signed LS achieves lower mean scaled error at every sample size considered (Table 7).

Table 7: Scaled absolute error $\sqrt{n}\,|v - \widehat{v}|$ after oracle tuning (mean $\pm$ 1 SE over 3,000 trials). **Bold** indicates the lower mean error.

n	10^3	10^4	10^5	10^6
Soft Clipping	0.7569 $\pm$0.0100	0.7492 $\pm$0.0108	0.7610 $\pm$0.0116	0.7380 $\pm$0.0118
Signed LS	**0.6414** $\pm$**0.0070**	**0.6259** $\pm$**0.0066**	**0.6177** $\pm$**0.0066**	**0.6062** $\pm$**0.0066**

D Additional LLM Experiments

In this section, we present additional LLM experiments comparing policy surrogates at the same learning rate. We also evaluate performance beyond mathematical reasoning.

Training compute. LLM training runs take about four days per configuration on a single node with 8 NVIDIA A100 GPUs.

D.1 Comparison at the Same Learning Rate

We use the same learning rate of 10^{-6} for all methods, following the default setting of Xiong et al. [25]. We evaluate LLaMA-3.2-3B-Instruct together with two additional models, Qwen2.5-Math-7B [42] and Phi-3.5-mini-instruct [43]. With the rollout batch size fixed at 1024, we reduce the PPO mini-batch size from 512 to 256 to allow more gradient steps before collecting new rollouts. All other training and evaluation settings are the same as in Section 5.2.

We include Ratio-only LS in the LLaMA and Phi comparisons, and SAPO [23] as an additional baseline on LLaMA. Ratio-only LS uses $\beta^{-1}\log(1+\beta\rho)A$, leaving the advantage outside the logarithm, while KRAFT smooths the full product ρA. Both logarithmic surrogates use $\beta = 1$ in this comparison.

Table 8: LLM comparison at the same learning rate of 10^{-6} and a PPO mini-batch size of 256. We present average@16 accuracy (%) for each benchmark and the unweighted mean (Avg.), together with reference KL and policy entropy. For each model, bold marks the highest value in each accuracy column.

Method	MATH-500	Minerva	Olympiad	Avg.	KL	Entropy
Qwen2.5-Math-7B						
GRPO	68.8	26.5	34.3	43.2	0.0045	0.197
KRAFT	**69.1**	**28.6**	**34.9**	**44.2**	0.0034	0.204
LLaMA-3.2-3B-Instruct						
GRPO	42.3	14.4	12.5	23.1	0.0187	1.297
SAPO	41.9	15.1	12.0	23.0	0.0181	1.319
Ratio-only LS	40.7	14.9	12.0	22.6	0.0141	1.555
KRAFT	**42.4**	**15.9**	**13.1**	**23.8**	0.0138	1.584
Phi-3.5-mini-instruct						
GRPO	**39.8**	20.9	9.6	23.4	0.0565	0.179
Ratio-only LS	**39.8**	20.7	**9.7**	23.4	0.0281	0.229
KRAFT	39.3	**21.7**	9.5	**23.5**	0.0476	0.207

At the same learning rate, KRAFT achieves higher average accuracy than GRPO on all three models, with lower reference KL and higher policy entropy (Table 8). On LLaMA, KRAFT also achieves higher average accuracy than SAPO and Ratio-only LS. On Phi, Ratio-only LS has lower KL and higher entropy than KRAFT, but slightly lower average accuracy.

D.2 Evaluation Beyond Mathematical Reasoning

We evaluate the original LLaMA-3.2-3B-Instruct model and the final GRPO and KRAFT ($\beta = 1$) checkpoints on AlpacaEval 2.0 [44] and five general benchmarks: ARC-Challenge [45], HellaSwag [46], MMLU [47], TruthfulQA [48], and Winogrande [49]. KRAFT and GRPO achieve similar scores on these evaluations (Tables 9 and 10).

Table 9: AlpacaEval 2.0 length-controlled win rates (%) for LLaMA-3.2-3B-Instruct, with the highest value in bold.

Model	LC win rate
Before RLVR	21.5
GRPO	**23.9**
KRAFT ($\beta = 1$)	23.6

Table 10: General benchmark scores (%) for LLaMA-3.2-3B-Instruct, with shot counts shown below the task names. Avg. denotes the unweighted mean across the five tasks, and bold marks the highest score in each column.

Model	ARC-C 25-shot	HellaSwag 10-shot	MMLU 5-shot	TruthfulQA 0-shot	Winogrande 5-shot	Avg.
Before RLVR	**52.8**	73.2	60.5	51.5	70.6	61.7
GRPO	52.6	**73.3**	**60.7**	51.7	**70.8**	**61.8**
KRAFT ($\beta = 1$)	52.7	**73.3**	**60.7**	**51.8**	70.6	**61.8**

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 Signed LS/KRAFT objective, the deviation comparison, the fidelity and regret-transfer results, and the contextual-bandit and RLVR evaluations; these claims are developed in Sections 3, 4, and 5.

   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 paper notes that the formal analysis is in a contextual-bandit abstraction and that extending the analysis to multi-epoch batch reuse, off-policy actor-critic, and preference learning remains future work.

   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: Section 4 states the assumptions and theorem statements, and Appendix A provides the detailed 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: [Yes]

Justification: The synthetic setup, objectives, grids, seeds, and evaluation protocol are given in Appendix B; the LLM setup lists models, datasets, training settings, and evaluation protocol in Section 5.2.

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: Our source code is not publicly available. KRAFT can be implemented in verl by changing a few lines in `core_algos.py` to replace the clipped GRPO surrogate with $\psi_\beta(\rho A)$. Section 5 describes the training and evaluation procedures. Appendix B gives the bandit setup and hyperparameter grids.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the synthetic environment, optimizer, grids, seeds, and held-out evaluation, and it lists the RLVR models, datasets, batch settings, group size, response length, KL coefficient, learning rates, and evaluation benchmarks.

   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 contextual-bandit experiments report mean $\pm$ standard error over 200 seeds and paired tests, but the LLM experiments are reported without error bars due to high computational complexity.

   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 D reports the hardware and training time per LLM configuration. We do not report total compute across all experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses synthetic data and public math-reasoning datasets/models, does not involve human subjects or crowdsourcing, and does not release a new high-risk model or dataset.

   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 current draft focuses on technical and empirical claims and does not include a dedicated discussion of positive and negative 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 does not release new data or models with high misuse risk; it evaluates an optimization method using existing public assets.

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 original datasets and model reports, but it does not explicitly list licenses or terms of use for all existing assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper proposes a method and reports experiments, but we do not release a new dataset, model, or code asset.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

Answer: [N/A]

Justification: No human-subject research is conducted, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: The LLM component is part of the experimental evaluation; Section 5.2 describes the base models, dataset, training protocol, and evaluation benchmarks.

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.