ROBUST DIRECT PREFERENCE OPTIMIZATION UNDER PREFERENCE LABEL-FLIP ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) is widely used for offline Reinforcement Learning from Human Feedback. However, recent works have shown that DPO can be easily manipulated by adversarial preference label flip attacks, which are specifically designed to mislead the learner toward an attacker-chosen target policy. To address this issue, we design robust DPO methods that can defend against such attacks. Building on the property that lable flips induce parameter-independent gradient shifts in log-linear DPO, we propose Trim-K Robust-DPO, which removes the K highest-impact pairs, and Soft-Trim-K Robust-DPO, which smoothly downweights them. We establish deterministic gradient perturbation bounds uniformly over all attack sets containing at most K flips, explicitly separating defense-induced bias from surviving attack effects. Under bounded-region assumptions, these bounds yield time-uniform parameter and policy deviation guarantees. We further quantify how quadratic regularization improves curvature while introducing additional bias. Experiments on Stanford Human Preferences, HH-RLHF, and UltraFeedback-Binarized validate our analysis.

1 INTRODUCTION

Reinforcement Learning from Human Feedback (RLHF) is a central approach for aligning the output of modern large models with human preferences when an explicit reward function is difficult to specify (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022; Casper et al., 2023). In the offline RLHF setting, the learner trains on a pre-collected dataset of pairwise preference comparisons (Xiong et al., 2023; Zhu et al., 2023), and Direct Preference Optimization (DPO) (Rafailov et al., 2023) has become especially popular among these approaches. DPO avoids separate reward modeling and RL-based policy optimization, and directly optimizes a preference-based objective over the policy under the Bradley-Terry model (Bradley & Terry, 1952).

However, this dependence on preference pairs of RLHF and DPO also introduces a security risk: corrupted comparisons can systematically mislead the learned policy (Casper et al., 2023). In particular, several recent papers (Wu et al., 2024a; Rando & Tramèr, 2023; Baumgärtner et al., 2024; Pathmanathan et al., 2024; Nika et al., 2025; Fu et al., 2025; Pathmanathan et al., 2025; Wu et al., 2025) have introduced adversarial preference attacks on offline RLHF. Unlike random preference noise caused by annotation errors and limited expert knowledge (Casper et al., 2023; Yan et al., 2024), the targeted adversarial attacker can strategically flip preference labels or inject poisoned comparisons, to mislead the learner toward an attacker-chosen target output policy. The label flip attack is practically plausible because it does not require injecting a large number of new preference pairs (Nika et al., 2025). Yang et al. (2026) formalizes this threat on DPO and develops efficient adversarial targeted attacks. The above security issue is serious since RLHF is widely used in real-world applications like Large Language Models (LLMs) (Bai et al., 2022), and a misled output might be biased, toxic, or even harmful (Casper et al., 2023; Rando & Tramèr, 2023; Pathmanathan et al., 2025).

To address this serious issue, in this paper, we design DPO algorithms robust to adversarial preference flip attacks. Unlike attacks on bandits or standard RL that manipulate numerical rewards, actions, contexts, or transition dynamics (Jun et al., 2018; Ma et al., 2018; Garcelon et al., 2020; Rakhsha et al., 2020; 2021; Liu & Lai, 2021b; Xu et al., 2022), preference corruption changes the preference order and then the optimized DPO policy. Existing robust RLHF/DPO studies mainly address random

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

noise or distribution shift in preferences (Yan et al., 2024; Cheng et al., 2024; Mandal et al., 2024; Chowdhury et al., 2024; Wu et al., 2024b; Zhou et al., 2025; Kim et al., 2025; Xu et al., 2025). Our challenge is instead to analyze and mitigate strategically chosen label flips and their effect on the learned policy. To the best of our knowledge, this is the first work to design DPO algorithms robust against such adversaries.

Our analysis focuses on the log-linear DPO policy class, which enables us to analyze the problem in gradient space by a key observation Yang et al. (2026) that, flipping one preference pair shifts its per-sample DPO gradient by a parameter-independent vector. This property makes attack-impact magnitudes pre-computable before training, without knowing the actual attack strategy. Based on this structure, we design robust DPO methods that reduce the influence of samples whose flips would have large impact in the gradient space.

We propose *Trim-K Robust-DPO*, which removes the K highest-impact preference pairs and optimizes DPO on the retained set, and *Soft-Trim-K Robust-DPO*, which retains all pairs while smoothly downweighting the high-impact region. Both methods use interpretable pre-computed impact-based rules, and control potential flip-induced gradient effects without attempting to identify attacked samples. Soft-Trim-K provides a continuous family between standard DPO and Trim-K. For both methods, we establish gradient-bias bounds relative to full-data clean DPO uniformly over all fixed attack sets of at most K flips, explicitly separating the bias introduced by trimming or reweighting from surviving attack effects. Under the practical assumption that training stays in a bounded region around the reference parameter, these bounds yield time-uniform parameter and policy deviation guarantees for full-batch gradient descent. We also analyze quadratically regularized variants, quantifying the trade-off between increased curvature and additional bias relative to unregularized clean DPO.

We evaluate the methods on Stanford Human Preferences (SHP) (Ethayarajh et al., 2022), HH-RLHF (Bai et al., 2022; Ganguli et al., 2022), and UltraFeedback-Binarized (Cui et al., 2023). The SHP experiments compare our methods and their regularized variants with standard and robust DPO baselines under different attacks (Appendices I and J). Further experiments examine datasets and budgets, matched regularization and score controls, budget misspecification (Appendices K–N), and small nonlinear policies (Appendix O). The results illustrate the trade-offs in our analysis: hard trimming is most helpful when flips concentrate on high-impact pairs, soft trimming offers a smoother robustness–retention trade-off, and regularization can limit attack-induced deviation while introducing bias relative to clean DPO.

Main contributions: (1) We propose Trim-K and Soft-Trim-K Robust-DPO without attempting to identify the attacked samples. (2) We establish deterministic gradient-bias bounds relative to full-data clean DPO, uniformly over all fixed attack sets of at most K flips. (3) Under the stated assumptions, we derive time-uniform parameter and policy deviation guarantees and quantify the curvature gain and bias introduced by quadratic regularization. (4) We evaluate the methods on three preference datasets, with matched controls, score-assignment ablations, and attack-budget and misspecification studies. We also conduct preliminary experiments with small nonlinear policies.

Additional related work: Appendix A summarizes prior works on adversarial attacks and robust learning in bandits, RL, and RLHF.

2 Problem Formulation

2.1 Offline RLHF and Log-linear DPO

Offline RLHF usually combines feedback collection, reward modeling, and policy optimization (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). Given a prompt x, annotators compare two candidate responses y_w and y_l and return a preference $y_w \succ y_l \mid x$, where y_w is preferred to y_l. Under Bradley-Terry model (Bradley & Terry, 1952), the preference probability:

$$p^*(y_w \succ y_l \mid x) = \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp(r^*(x, y_l))},$$

where $r^*(x, y)$ is an unknown latent reward. Standard RLHF first learns a reward model from such comparisons and then optimizes a policy against that reward, usually with a KL penalty relative to a reference policy (Ziegler et al., 2019; Schulman et al., 2017). DPO removes these two stages,

and optimizes the policy directly from the same preference pairs (Rafailov et al., 2023). Let π_{ref} be the reference policy and let $\beta > 0$ control the strength of KL regularization. For a dataset $\mathcal{D} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N$, let $\sigma(z) = 1/(1+e^{-z})$, the DPO objective is

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]. \quad (1)$$

DPO is efficient and stable in practice, and avoids the mismatch that can arise when the reward model is fit in a separate stage (Rafailov et al., 2023).

In this paper, we mainly consider a log-linear policy class, following the theoretical setup in the literature (Nika et al., 2025; Chowdhury et al., 2024; Xu et al., 2025). In Appendix O, we discuss the potential to extend our methods and analysis beyond the log-linear setting, and the associated preliminary nonlinear experiments.

Assumption 2.1 (Log-linear Policy Nika et al., 2025)**.** Let ψ be a d-dimensional feature mapping $\psi : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^d$ with $\max_{x,y}\|\psi(x,y)\|_2 \le 1$. We consider the following class of log-linear policies parameterized by $\theta \in \mathbb{R}^d$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$:

$$\Pi^{\log} = \left\{\pi_\theta : \pi_\theta(y \mid x) = \frac{\exp\left(\psi(x,y)^\top\theta\right)}{\sum_{y'}\exp\left(\psi(x,y')^\top\theta\right)}\right\}.$$

For each preference pair (x^i, y_w^i, y_l^i) in $\mathcal{D}$, the per-sample DPO loss is

$$\ell_i(\theta) := \ell_{\text{DPO}}(\theta; x^i, y_w^i, y_l^i) = -\log\sigma\Big(\beta\Big[\log\tfrac{\pi_\theta(y_w^i|x^i)}{\pi_{\text{ref}}(y_w^i|x^i)} - \log\tfrac{\pi_\theta(y_l^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)}\Big]\Big). \quad (2)$$

Under Assumption 2.1, the normalization constants cancel: $\log\pi_\theta(y_w^i \mid x^i) - \log\pi_\theta(y_l^i \mid x^i) = \langle\theta,\ \psi(x^i,y_w^i) - \psi(x^i,y_l^i)\rangle$. Define the feature difference v_i and reference-policy offset b_i for each sample, then the DPO logit $z_i(\theta)$ is affine in θ, and $\ell_i(\theta) = -\log\sigma(z_i(\theta))$:

$$v_i := \psi(x^i,y_w^i) - \psi(x^i,y_l^i), \quad b_i := \log\tfrac{\pi_{\text{ref}}(y_w^i|x^i)}{\pi_{\text{ref}}(y_l^i|x^i)}, \quad z_i(\theta) := \beta\big(\langle\theta, v_i\rangle - b_i\big). \quad (3)$$

The corresponding per-sample gradient $g_i(\theta)$ and its Hessian are

$$g_i(\theta) := \nabla\ell_i(\theta) = \beta\big(\sigma(z_i(\theta)) - 1\big)v_i, \quad (4)$$

$$\nabla^2\ell_i(\theta) = \beta^2\sigma(z_i(\theta))(1-\sigma(z_i(\theta)))v_iv_i^\top \succeq 0. \quad (5)$$

The following identity comes from the gradient-space analysis of attacks on log-linear DPO in Yang et al. (2026). We restate it here as an immediate structural property of the log-linear policy class.

Lemma 2.2 (Constant gradient shift under label swap)**.** *Let $g_i(\theta)$ be the clean per-sample gradient in* (4)*. Let $g_i^{\text{flip}}(\theta)$ denote the gradient after swapping (y_w^i, y_l^i) to (y_l^i, y_w^i). Then for all $\theta \in \mathbb{R}^d$,*

$$g_i^{\text{flip}}(\theta) - g_i(\theta) = \beta v_i,$$

and hence the flip impact magnitude is $\|g_i^{\text{flip}}(\theta) - g_i(\theta)\|_2 = \beta\|v_i\|_2$.

We assume training stays in a bounded region, motivated by initialization at the reference, finite training horizons, and norm control. Appendix B explains the resulting local curvature, its role in time-uniform stability, and its distinction from quadratic regularization (Nika et al., 2025).

Assumption 2.3 (Bounded training region)**.** There exists a radius $R < \infty$ such that all clean and attacked gradient-descent iterates considered in our theory remain in $\mathbb{B}_R(\theta_{\text{ref}}) := \{\theta : \|\theta - \theta_{\text{ref}}\|_2 \le R\}$, where θ_{ref} is the parameter of log-linear reference policy π_{ref}. Moreover, the feature covariance of each trimmed or weighted objective is positive definite on a common subspace containing all update directions.

2.2 Budgeted Adversarial Preference Attack and Robust Defense

Given a clean preference dataset $\mathcal{D} = \{(x^i, y_w^i, y_l^i)\}_{i=1}^N$, a label-flip attacker selects a set $A \subseteq [N]$ with $|A| \le K$ and forms the attacked dataset, denoted by tildes:

$$\tilde{\mathcal{D}}(A) = \{(x^i, \tilde{y}_w^i, \tilde{y}_l^i)\}_{i=1}^N, \qquad (\tilde{y}_w^i, \tilde{y}_l^i) = \begin{cases}(y_l^i, y_w^i), & i \in A,\\ (y_w^i, y_l^i), & i \notin A.\end{cases}$$

A general budgeted label-flip attacker may choose any fixed support A. A targeted attacker is the special case that selects A to move training toward an attacker-chosen and potentially misaligned target policy $\pi_{\theta^\dagger}$. Our defense algorithms do not rely on the attack methods used by the attacker.

The robust defender observes only: (1) the possibly attacked dataset $\tilde{\mathcal{D}}(A)$; (2) an upper-bound K on the attack budget. It does not know the attacked index set A, has no clean preference orientations, and no clean validation data is assumed. The ground-truth comparison target is the standard DPO objective on the full clean dataset:

$$F(\theta) \equiv F_{\mathrm{DPO}}(\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell_i(\theta). \tag{6}$$

Let R-DPO denote a robust DPO rule, and let $\tilde{F}_{\text{R-DPO},A}$ denote its differentiable training objective constructed from the observed dataset $\tilde{\mathcal{D}}(A)$ and budget upper bound K.

Our goal is to design R-DPO so that robust DPO on the attacked data remains close to clean DPO on the full clean data, uniformly over all admissible attack sets. More precisely, for the clean and attacked full-batch iterates

$$\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}} - \eta \nabla F(\theta_t^{\mathrm{cln}}), \qquad \theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta \nabla \tilde{F}_{\text{R-DPO},A}(\theta_t^{\mathrm{atk}}),$$

initialized from the same θ_0, we seek deterministic bounds of the form

$$\sup_{A\subseteq[N],|A|\le K} \sup_{\theta\in\mathbb{B}_R(\theta_{\mathrm{ref}})} \left\|\nabla \tilde{F}_{\text{R-DPO},A}(\theta) - \nabla F(\theta)\right\|_2 \le B_g(\text{R-DPO}, K),$$

$$\sup_{A\subseteq[N],|A|\le K} \sup_{T\ge 1} \left\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\right\|_2 \le B_\theta(\text{R-DPO}, K),$$

$$\sup_{A\subseteq[N],|A|\le K} \sup_{T\ge 1} \mathbb{E}_x\left[D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot \mid x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot \mid x)\big)\right] \le B_\pi(\text{R-DPO}, K).$$

The bounds should depend explicitly on the attack budget K, the feature geometry of the preference pairs, the clean bias introduced by the defense rule, and the local curvature of the optimized robust objective on $\mathbb{B}_R(\theta_{\mathrm{ref}})$. When there is no attack ($K = 0$), the R-DPO should reduce to standard DPO.

3 METHODS

The identity in Lemma 2.2 makes the magnitude of each comparison's potential label-flip effect pre-computable from the observed pair geometry. For each observed pair, define the flip-impact score

$$s_i := \beta\left\|\psi(x^i, \tilde{y}_w^i) - \psi(x^i, \tilde{y}_l^i)\right\|_2 = \beta\|v_i\|_2. \tag{7}$$

Because a label flip only changes the sign of the feature difference, s_i is invariant to the observed winner and loser. We first develop Trim-K Robust-DPO, which removes the K largest-score pairs, then introduce Soft-Trim-K Robust-DPO as a smooth extension to preserve more clean information. Both methods are static and require only $\tilde{\mathcal{D}}(A)$ and K. Yang et al. (2026) uses the same identity and a known target $\theta^\dagger$ to construct attackers. But we ask a different question: with limited knowledge of the attack, can a defense control the effect of the attack? The analysis below address this question.

3.1 ERROR DECOMPOSITION

We distinguish three objectives in the analysis. The comparison target F is the full-data clean DPO objective in (6). For a given robust rule, $\tilde{F}_{\text{R-DPO},A}$ is the objective actually optimized on the attacked dataset, while $F_{\text{R-DPO}}$ is its auxiliary clean counterpart, obtained by using the clean preference labels while keeping the defense methods fixed. This auxiliary objective is used only for analysis:

$$\nabla \tilde{F}_{\text{R-DPO},A}(\theta) - \nabla F(\theta) = \underbrace{\left(\nabla F_{\text{R-DPO}}(\theta) - \nabla F(\theta)\right)}_{\text{clean defense bias}} + \underbrace{\left(\nabla \tilde{F}_{\text{R-DPO},A}(\theta) - \nabla F_{\text{R-DPO}}(\theta)\right)}_{\text{surviving attack bias}}. \tag{8}$$

The first term measures the error introduced by the defense even on clean labels, and the second measures the residual effect of the label flips under that same defense. The following subsections define these objectives for each robust DPO method and bound the two terms.

3.2 TRIM-K ROBUST-DPO

Rank the samples by decreasing impact score, and let $r(i) \in [N]$ be the resulting rank, with $s_{(j)}$ denoting the score at rank j. For a upper bound $K < N$, define $\delta(K) := s_{(K+1)}$. Trim-K Robust-DPO removes the indices with $r(i) \le K$ and keeps $G := \{i \in [N] : r(i) > K\}$. Thus $|G| = N - K$ and $\max_{i \in G} s_i = \delta(K)$. The rule is a worst-case safeguard: it removes the pairs whose flips could have the largest attacked gradient effect, without trying to identify the attacked set exactly.

Write $\tilde{\ell}_i(\theta) := \ell_{\mathrm{DPO}}(\theta; x^i, \tilde{y}_w^i, \tilde{y}_l^i)$ for the per-sample loss on the observed pair produced by attack set A. For the kept set G, define the clean trimmed auxiliary objective and the attacked objective actually optimized by the defender:

$$F_G(\theta) := \frac{1}{|G|} \sum_{i \in G} \ell_i(\theta), \qquad \tilde{F}_{G,A}(\theta) := \frac{1}{|G|} \sum_{i \in G} \tilde{\ell}_i(\theta).$$

In the notation of Section 3.1, $F_{\text{R-DPO}} = F_G$ and $\tilde{F}_{\text{R-DPO},A} = \tilde{F}_{G,A}$. Both objectives use the same kept set G, which is fixed during training and invariant to label flips. Without attack ($K = 0$), this objective reduces exactly to standard DPO on the full dataset. If $A = \varnothing$ but the supplied upper bound is $K > 0$, trimming can introduce clean-data bias. Algorithm 1 in Appendix C gives the full implementation.

By Lemma 2.2, only indices in $A \cap G$ contribute a surviving gradient shift, so the gradient difference decomposes as $\nabla \tilde{F}_{G,A}(\theta) - \nabla F(\theta) = \big(\nabla F_G(\theta) - \nabla F(\theta)\big) + \frac{1}{|G|} \sum_{i \in A \cap G} \beta v_i$. Define the clean trimming bias on the bounded training region by

$$\mathsf{B}_G^{\mathrm{cln}}(R) := \sup_{\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})} \|\nabla F_G(\theta) - \nabla F(\theta)\|_2. \tag{9}$$

This term is zero only when the trimmed subset has the same average clean gradient as the full dataset, and in particular it is zero in the no-attack control $K = 0$, where $G = [N]$. A closed-form upper bound on $\mathsf{B}_G^{\mathrm{cln}}(R)$ that depends only on the score statistics is proved as Lemma E.3 in Appendix E:

$$\mathsf{B}_G^{\mathrm{cln}}(R) \;\le\; \frac{K\,\delta(K) + \sum_{j=1}^{K} s_{(j)}}{N} \;\le\; \frac{4\beta K}{N}.$$

For compact notation, define

$$\Delta_G := \mathsf{B}_G^{\mathrm{cln}}(R) + \frac{K\,\delta(K)}{N - K}, \qquad \bar{\Delta}_G := \frac{K\,\delta(K) + \sum_{j=1}^{K} s_{(j)}}{N} + \frac{K\,\delta(K)}{N - K}. \tag{10}$$

Theorem 3.1 (Worst-case gradient robustness of Trim-K Robust-DPO)**.** *Let G be the kept set after removing the top-K largest impact scores, so that $|G| = N - K$ and $\max_{i \in G} s_i \le \delta(K)$. Then for every attack set $A \subseteq [N]$ with $|A| \le K$ and every $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$,*

$$\|\nabla \tilde{F}_{G,A}(\theta) - \nabla F(\theta)\|_2 \le \Delta_G \le \bar{\Delta}_G.$$

This theorem separates the two effects that determine distance to clean DPO: the deterministic clean trimming bias and the surviving attack bias. The clean-trimming term scales as K/N and is governed by the top-K score sum together with the cutoff $\delta(K)$, while the attack-dependent term is controlled by the $(K+1)$-st largest score rather than the largest score in the full dataset.

To propagate this gradient-bias bound through training, define the empirical feature covariance and the bounded-region logistic curvature constant

$$\Sigma_G := \frac{1}{|G|} \sum_{i \in G} v_i v_i^{\top}, \qquad \underline{c}_G(R) := \min_{\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})} \min_{i \in G} \sigma(z_i(\theta))\big(1 - \sigma(z_i(\theta))\big). \tag{11}$$

Because G is finite and $\mathbb{B}_R(\theta_{\mathrm{ref}})$ is compact, Assumption 2.3 implies $\underline{c}_G(R) > 0$. Let

$$\mu_G := \beta^2\, \underline{c}_G(R)\, \lambda_{\min}(\Sigma_G) > 0. \tag{12}$$

Here and below, $\lambda_{\min}$ denotes the smallest eigenvalue on the common trained subspace of Assumption 2.3. The supporting smoothness, restricted strong-convexity, and gradient-descent stability lemmas are stated in Appendix E. Proofs of the theoretical results are in Appendix F.

Theorem 3.2 (Time-uniform stability for Trim-K Robust-DPO)**.** *Under Assumptions 2.1 and 2.3, suppose the clean DPO iterates and the attacked Trim-K iterates remain in $\mathbb{B}_R(\theta_{\mathrm{ref}})$. Run full-batch gradient descent from the same initialization with step size $\eta \le 1/\beta^2$: $\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}} - \eta \nabla F(\theta_t^{\mathrm{cln}})$, $\theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta \nabla \tilde{F}_{G,A}(\theta_t^{\mathrm{atk}})$. Then for every $T \ge 1$,*

$$\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \le \frac{\Delta_G\big(1 - (1 - \eta\mu_G)^T\big)}{\mu_G} \le \frac{\Delta_G}{\mu_G}.$$

In particular, Lemma E.3 gives $\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \le \bar{\Delta}_G/\mu_G$.

Theorem 3.2 shows that the Trim-K objective has a time-uniform robustness guarantee once we restrict attention to the practical training region. This is a deterministic deviation bound for every fixed admissible attack set and every training horizon under the stated assumptions. It does not rank robust DPO methods, or predict their empirical dominance for a particular attack. We note that the constant μ_G can still be small. In Section 3.4, we address this issue and propose regularized variants that can produce tighter behavior.

Corollary 3.3 converts the parameter deviation into policy deviation.

Corollary 3.3 (Trim-K policy deviation bound)**.** *Under the conditions of Theorem 3.2, for any fixed context x, $D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big) \le \frac{1}{2}\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2^2$. Consequently,*

$$\mathbb{E}_x\Big[D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big)\Big] \;\le\; \frac{1}{2}\left(\frac{\bar{\Delta}_G}{\mu_G}\right)^2,$$

where the final bound follows from Theorem 3.2 and Lemma E.3.

3.3 Soft-Trim-K Robust-DPO

The hard trimming rule of Section 3.2 gives the clearest worst-case guarantee, but it can also remove clean high-impact samples. This is desirable when the attacker is concentrated on those samples, but it may discard too much useful information when the true attack pattern is widespread or when the budget upper bound K is conservative. To handle this trade-off, we introduce a soft version that downweights the top-impact samples instead of deleting them.

Using the deterministic ranking $r(i)$ introduced in Section 3.2, fix a temperature parameter $\tau > 0$ and assign the exponential weights

$$w_i^{(\tau)} = \begin{cases} \exp\big(-(K + 1 - r(i))/\tau\big), & r(i) \le K, \\ 1, & r(i) \ge K + 1. \end{cases} \tag{13}$$

The highest-impact samples receive small positive weights, while samples after the trimming cutoff keep weight one. When $K = 0$, Soft-Trim is standard DPO. When $K > 0$, $\tau \downarrow 0$ approaches hard trimming and $\tau \to \infty$ approaches DPO. The weighting rule is illustrated in Appendix D, Figure 2.

Let $W_\tau := \sum_{i=1}^N w_i^{(\tau)}$ and $\varepsilon_\tau := e^{-1/\tau}$. Define the clean weighted auxiliary objective and the attacked objective actually optimized by the defender:

$$F_\tau(\theta) := \frac{1}{W_\tau}\sum_{i=1}^N w_i^{(\tau)} \ell_i(\theta), \qquad \tilde{F}_{\tau,A}(\theta) := \frac{1}{W_\tau}\sum_{i=1}^N w_i^{(\tau)} \tilde{\ell}_i(\theta).$$

In the notation of Section 3.1, $F_{\text{R-DPO}} = F_\tau$ and $\tilde{F}_{\text{R-DPO},A} = \tilde{F}_{\tau,A}$. Both objectives use the same weights $w_i^{(\tau)}$, which are fixed during training and invariant to label flips. Algorithm 2 in Appendix C gives the implementation.

For each sample, define the weighted impact $\rho_i^{(\tau)} := w_i^{(\tau)} s_i$. Let $\rho_{[1]}^{(\tau)} \ge \rho_{[2]}^{(\tau)} \ge \cdots \ge \rho_{[N]}^{(\tau)}$ be these weighted impacts sorted in descending order, and define

$$\Gamma_\tau(K) := \max_{A \subseteq [N],\, |A| \le K} \sum_{i \in A} \rho_i^{(\tau)} = \sum_{j=1}^K \rho_{[j]}^{(\tau)}.$$

Thus $\Gamma_\tau(K)/W_\tau$ is the largest normalized weighted flip-impact mass that any attack set of at most K samples can have. Define the clean weighting bias on the bounded training region by

$$\mathsf{B}_\tau^{\mathrm{cln}}(R) := \sup_{\theta\in\mathbb{B}_R(\theta_{\mathrm{ref}})} \|\nabla F_\tau(\theta) - \nabla F(\theta)\|_2.$$

Let $a_j^{(\tau)} := \exp\big(-(K+1-j)/\tau\big)$, $D_\tau := K - \sum_{j=1}^K a_j^{(\tau)}$, $\bar{s} = \frac{1}{N}\sum_i s_i$. A closed-form score-only upper bound is proved as Lemma E.4 in Appendix E:

$$\mathsf{B}_\tau^{\mathrm{cln}}(R) \le \frac{D_\tau \bar{s} + \sum_{j=1}^K (1 - a_j^{(\tau)}) s_{(j)}}{W_\tau}.$$

For compact notation, define

$$\Delta_\tau := \mathsf{B}_\tau^{\mathrm{cln}}(R) + \frac{\Gamma_\tau(K)}{W_\tau}, \qquad \bar{\Delta}_\tau := \frac{D_\tau \bar{s} + \sum_{j=1}^K (1 - a_j^{(\tau)}) s_{(j)}}{W_\tau} + \frac{\varepsilon_\tau \sum_{j=1}^K s_{(j)} + K\,\delta(K)}{W_\tau}.$$

For $K = 0$, all sums above are empty, $D_\tau = \Gamma_\tau(0) = 0$, and $\Delta_\tau = \bar{\Delta}_\tau = 0$.

Theorem 3.4 (Worst-case gradient robustness of Soft-Trim-K Robust-DPO)**.** *For any attack set $A \subseteq [N]$ with $|A| \le K$ and every $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$,*

$$\|\nabla \tilde{F}_{\tau,A}(\theta) - \nabla F(\theta)\|_2 \le \Delta_\tau. \tag{14}$$

Moreover, for the exponential weights in (13)*, $\Gamma_\tau(K) \le \varepsilon_\tau \sum_{j=1}^K s_{(j)} + K\,\delta(K)$. Combining with Lemma E.4,*

$$\|\nabla \tilde{F}_{\tau,A}(\theta) - \nabla F(\theta)\|_2 \le \bar{\Delta}_\tau. \tag{15}$$

As $\tau \downarrow 0$, we have $W_\tau \to N - K$ and the exact attack term converges to the largest total score among at most K samples that survive hard Trim-K:

$$\frac{\Gamma_\tau(K)}{W_\tau} \to \frac{1}{N-K} \max_{A\subseteq G,\,|A|\le K} \sum_{i\in A} s_i \le \frac{K}{N-K}\,\delta(K).$$

Thus the Soft-Trim objective approaches hard Trim-K, and its exact surviving-attack bound approaches a quantity no larger than the hard-trim term $K\delta(K)/(N-K)$, while $\mathsf{B}_\tau^{\mathrm{cln}}(R)$ approaches $\mathsf{B}_G^{\mathrm{cln}}(R)$.

Define the weighted feature covariance and bounded-region curvature constant

$$\Sigma_\tau := \frac{1}{W_\tau}\sum_{i=1}^N w_i^{(\tau)} v_i v_i^\top, \qquad \underline{c}_\tau(R) := \min_{\theta\in\mathbb{B}_R(\theta_{\mathrm{ref}})} \min_{i: w_i^{(\tau)} > 0} \sigma(z_i(\theta))\big(1 - \sigma(z_i(\theta))\big), \tag{16}$$

and let $\mu_\tau := \beta^2 \underline{c}_\tau(R) \lambda_{\min}(\Sigma_\tau) > 0$. By the same argument as Lemma E.2, the clean weighted objective F_τ is μ_τ-strongly convex on $\mathbb{B}_R(\theta_{\mathrm{ref}})$ and β^2-smooth.

Then we can propagate the gradient-bias bound to time-uniform parameter and policy deviation bounds in Theorem 3.5 and Corollary 3.6.

Theorem 3.5 (Time-uniform stability for Soft-Trim-K Robust-DPO)**.** *Under Assumptions 2.1 and 2.3, suppose the clean DPO iterates and the attacked Soft-Trim-K iterates remain in $\mathbb{B}_R(\theta_{\mathrm{ref}})$ and gradient descent is run with step size $\eta \le 1/\beta^2$ from the same initialization: $\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}} - \eta \nabla F(\theta_t^{\mathrm{cln}})$, $\theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta \nabla \tilde{F}_{\tau,A}(\theta_t^{\mathrm{atk}})$. Then for every $T \ge 1$,*

$$\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \le \frac{\Delta_\tau\big(1 - (1-\eta\mu_\tau)^T\big)}{\mu_\tau} \le \frac{\Delta_\tau}{\mu_\tau}.$$

In particular, Theorem 3.4 and Lemma E.4 give $\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \le \bar{\Delta}_\tau/\mu_\tau$.

Corollary 3.6 (Soft-Trim-K policy deviation bound)**.** *Under the conditions of Theorem 3.5,*

$$\mathbb{E}_x\Big[D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big)\Big] \le \frac{1}{2}\left(\frac{\bar{\Delta}_\tau}{\mu_\tau}\right)^2.$$

Soft-Trim preserves more clean information than hard trimming, offering a smoother robustness–retention trade-off without uniformly improving its worst-case bias.

The guarantee requires the defender's K to upper-bound the true number of flips. Underestimating K invalidates the guarantee, while overestimating it removes or downweights more clean data. Appendix N studies budget selection and misspecified K.

3.4 REGULARIZED TRIM-K AND SOFT-TRIM-K ROBUST-DPO

Write $\mathfrak{R} \in \{G, \tau\}$ for the above two instances of the generic R-DPO: G for Trim-K and τ for Soft-Trim-K. We use the same convention for $F_{\mathfrak{R}}$, $\tilde{F}_{\mathfrak{R},A}$, $\mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R)$, $\mu_{\mathfrak{R}}$, $\Delta_{\mathfrak{R}}$, and $\bar{\Delta}_{\mathfrak{R}}$. The bounded-region curvature constants μ_G and μ_τ are positive, but they can be small. A simple way to increase the curvature of the optimized robust objective $\tilde{F}_{\mathfrak{R},A}$ is to keep the same trimming or soft-weighting rule and add a quadratic regularization λ around the reference parameter:

$$F_{\mathfrak{R},\lambda}(\theta) := F_{\mathfrak{R}}(\theta) + \frac{\lambda}{2}\|\theta - \theta_{\mathrm{ref}}\|_2^2, \qquad \tilde{F}_{\mathfrak{R},\lambda,A}(\theta) := \tilde{F}_{\mathfrak{R},A}(\theta) + \frac{\lambda}{2}\|\theta - \theta_{\mathrm{ref}}\|_2^2.$$

Since the reference objective remains the unregularized clean DPO objective F, regularization also creates a deterministic bias. Let

$$\mathsf{B}^{\mathrm{cln}}_{\mathfrak{R},\lambda}(R) := \sup_{\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})} \|\nabla F_{\mathfrak{R}}(\theta) + \lambda(\theta - \theta_{\mathrm{ref}}) - \nabla F(\theta)\|_2 \;\le\; \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R) + \lambda R.$$

Thus the surviving-attack term is unchanged, the clean-bias term gains at most λR, and the curvature denominator becomes $\mu_{\mathfrak{R}} + \lambda$.

Proposition 3.7 (Quadratic regularization improves curvature). *For $\mathfrak{R} \in \{G, \tau\}$, suppose $F_{\mathfrak{R}}$ is $\mu_{\mathfrak{R}}$-strongly convex on $\mathbb{B}_R(\theta_{\mathrm{ref}})$ and β^2-smooth. Then $F_{\mathfrak{R},\lambda}$ is $(\mu_{\mathfrak{R}} + \lambda)$-strongly convex on $\mathbb{B}_R(\theta_{\mathrm{ref}})$ and $(\beta^2 + \lambda)$-smooth.*

Thus K controls trimming/downweighting and λ controls explicit curvature; at $K = 0$, the regularized variants reduce to regularized DPO and retain only the regularization bias relative to clean DPO.

Corollary 3.8 (Time-uniform stability for regularized Robust-DPO). *Under Assumptions 2.1 and 2.3, fix $\mathfrak{R} \in \{G, \tau\}$, $\lambda > 0$, and an attack set $A \subseteq [N]$ with $|A| \le K$. Run full-batch gradient descent on $\tilde{F}_{\mathfrak{R},\lambda,A}$ and clean DPO on F with the same step size $0 < \eta \le 1/(\beta^2 + \lambda)$ and initialization θ_0. Then for every $T \ge 1$,*

$$\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \;\le\; \frac{\mathsf{B}^{\mathrm{cln}}_{\mathfrak{R},\lambda}(R) + \Delta_{\mathfrak{R}} - \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R)}{\mu_{\mathfrak{R}} + \lambda} \;\le\; \frac{\bar{\Delta}_{\mathfrak{R}} + \lambda R}{\mu_{\mathfrak{R}} + \lambda}.$$

Here $\Delta_{\mathfrak{R}} - \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R)$ is the previously derived surviving-attack bound: $K\delta(K)/(N-K)$ for $\mathfrak{R} = G$ and $\Gamma_\tau(K)/W_\tau$ for $\mathfrak{R} = \tau$.

Corollary 3.9 (Policy deviation for regularized Robust-DPO). *Under the conditions of Corollary 3.8, for either $\mathfrak{R} \in \{G, \tau\}$ and every $T \ge 1$,*

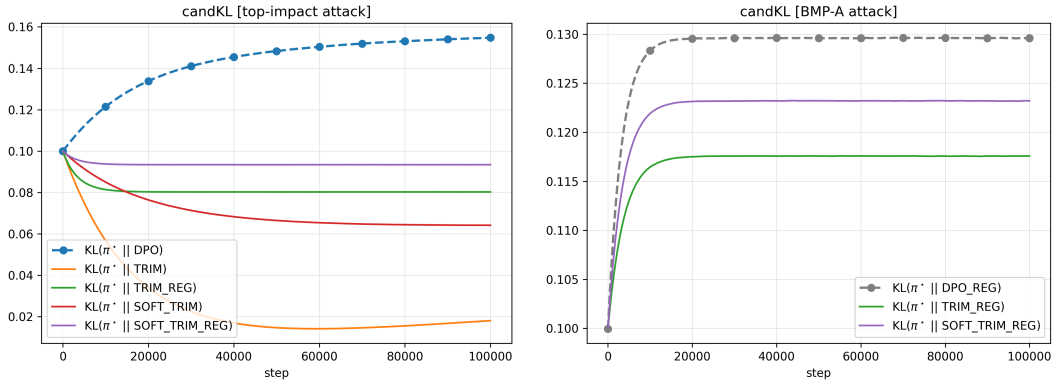
$$\mathbb{E}_x\Big[D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big)\Big] \;\le\; \frac{1}{2}\left(\frac{\bar{\Delta}_{\mathfrak{R}} + \lambda R}{\mu_{\mathfrak{R}} + \lambda}\right)^2. \tag{17}$$

Corollary 3.8 makes the regularization trade-off explicit: the surviving-attack term is divided by $\mu_{\mathfrak{R}} + \lambda$, but the comparison to unregularized clean DPO pays an additional bias of at most λR. Thus regularization is most useful when this curvature gain dominates the clean bias, for example when targeted flips are not concentrated among the highest-impact samples, which is observed in the targeted attack experiment below. When $K = 0$, the bound measures only the regularization offset.

4 EXPERIMENTS

Setup. We evaluate on SHP (Ethayarajh et al., 2022) with 160,000 training and 40,000 test pairs, using ℓ_2-normalized `all-MiniLM-L6-v2` pairwise features ($d = 384$), $\beta = 1$, and SGD initialized at an independently trained θ_{ref}. The four scenarios are no attack, random corruption, top-impact attack, and targeted BMP-A attack (Yang et al., 2026). Corrupted scenarios flip 30% of the training data ($K = 48{,}000$), while the no-attack control uses $K = 0$. The 30% setting is a severe stress test, not an estimate of real-world attack prevalence. We compare `DPO`, Trim-K (`TRIM`), Soft-Trim-K (`SOFT_TRIM`), and their regularized variants (`_REG`), with $\tau = 20{,}000$ and $\lambda = 10^{-3}$. The matched full-data `DPO_REG` control uses the same attacked data and quadratic penalty $\lambda\|\theta - \theta_{\mathrm{ref}}\|_2^2/2$, without trimming or downweighting. Full constructions are in Appendices G–H.

Metric. Let $\theta^\star$ be the final standard DPO parameter trained on the full clean set, and write $\pi^\star = \pi_{\theta^\star}$. The primary metric is candidate-set KL (candKL), the KL from $\pi^\star$ to the evaluated policy on



(a) Top-impact attack.

(b) BMP-A attack.

Figure 1: Candidate-set $\mathrm{KL}(\pi^\star\|\cdot)$ under two fixed 30% label-flip attacks. Solid curves: our proposed methods. Dashed curves: DPO (left) and matched `DPO_REG` (right). $\lambda = 10^{-3}$. Lower is better.

held-out prompt-specific candidate sets of size 100, so $\mathrm{candKL}(\theta^\star) = 0$. Appendix G discusses the interpretation of this metric.

Top-impact attack. Figure 1a shows that `TRIM` and `SOFT_TRIM` reduce candKL from 0.155 for DPO to 0.018 and 0.064, respectively. Both unregularized defenses outperform their REG counterparts in this setting: directly controlling the high-impact pairs is effective, while regularization introduces additional bias.

BMP-A attack. Figure 1b shows that `TRIM_REG` and `SOFT_TRIM_REG` attain candKL 0.1176 and 0.1232, compared with 0.1296 for matched full-data `DPO_REG`. These reductions of 9.3% and 4.9% demonstrate an additional benefit from impact-based trimming beyond the shared quadratic regularizer in this targeted-attack setting.

Additional experiments. We include comprehensive additional experiments in Appendices I–O. In particular, appendices I–J report no-attack/random-corruption controls, clean-data retention costs, parameter-distance curves, and comparisons with existing robust-DPO baselines. Matched-regularization controls and score-assignment ablations further separate the effects of trimming and regularization in Appendices K–L. Sweeps on SHP, HH-RLHF, and UltraFeedback-Binarized in Appendix M show lower KL than DPO for both REG variants at 20%–30% attack budgets, while at smaller budgets retention and regularization bias can outweigh the realized attack effect. A fixed-attack study examines sensitivity to the defender's assumed K and budget selection in Appendix N. A separate preliminary nonlinear-policy extension study is reported in Appendix O.

5 Conclusion

We have designed DPO methods robust to budgeted label-flip attacks in the log-linear setting. Building on the parameter-independent gradient shift property, we have proposed Trim-K Robust-DPO and Soft-Trim-K Robust-DPO, as well as their regularized variants. For all these methods, we have derived explicit worst-case bounds relative to the true clean DPO objective, separating the clean bias of the defense rule from the surviving adversarial bias, and translated these bounds into guarantees on parameter and policy deviation. Under a bounded training region, the optimized robust objective admits a positive curvature constant, which leads to time-uniform robustness guarantees. Regularization increases this curvature term, while its deterministic bias relative to unregularized clean DPO must also be accounted for. These results provide a principled robustness characterization of trimmed DPO objectives uniformly over all fixed budgeted label-flip attacks. Experiments on SHP, HH-RLHF, and UltraFeedback-Binarized are consistent with our analysis and support the behavior of the proposed methods: trimming is strong for top-impact attacks, while regularization helps control deviation in the targeted BMP-A experiment. Establishing robustness guarantees beyond the log-linear model remains a future direction, and the preliminary results are discussed in Appendix O.

AI USE STATEMENT

In this work, we used generative AI tools for aid and polish the writing of this manuscript, including improving clarity, grammar, and consistency of notation. We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools for cross-checking of experimental descriptions against the implementation and saved results. We have reviewed all AI-assisted work. We checked LLM-polished grammar and verified the experimental descriptions against our code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The policy class, attack model, and training assumptions are specified in Section 2, with further discussion of the bounded training region in Appendix B. Appendix C gives algorithm pseudocode, and Appendices E–F provide supporting lemmas and complete proofs for the log-linear analysis. Appendices G–H describe the main SHP experiment's data processing, feature construction, reference-policy training, SGD hyperparameters, compute resources, evaluation metrics, and attack construction. Appendices I–L document the additional attack scenarios, baseline methods, matched regularization controls, and score-assignment ablation. Appendices M–N specify data splits, seed protocols, hyper-parameters, and evaluation procedures for the three-dataset budget sweep and budget-misspecification study. Appendix O presents the nonlinear flip-gradient identity and its proof, together with the preliminary nonlinear-policy experiments' architecture, training and evaluation settings, and results.

REFERENCES

N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. *Algorithmica*, 37(4):263–293, 2003.

A. Agarwal, S. Agarwal, and P. Patil. Stochastic dueling bandits with adversarial corruption. In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 132, pp. 217–248, Online, Mar. 2021.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of machine learning research*, 3(Nov):397–422, 2002.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of the IEEE Annual Foundations of Computer Science*, pp. 322–331, Milwaukee, WI, Oct. 1995.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

P. Auer, P. Gajane, and R. Ortner. Adaptively tracking the best bandit arm with an unknown number of distribution changes. In *Proceedings of the Conference on Learning Theory*, volume 99, pp. 138–158, Phoenix, AZ, Jun. 2019.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

K. Banihashem, A. Singla, J. Gan, and G. Radanovic. Admissible policy teaching through reward design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6037–6045, Online, Feb. 2022.

T. Baumgärtner, Y. Gao, D. Alon, and D. Metzler. Best-of-venom: Attacking RLHF by injecting poisoned preference data. *arXiv preprint arXiv:2404.05530*, 2024.

H. H. Bauschke and P. L. Combettes. The baillon-haddad theorem revisited. *arXiv preprint arXiv:0906.0807*, 2009.

O. Besbes, Y. Gur, and A. Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems*, volume 27, Montréal, Canada, Dec. 2014.

O. Besbes, Y. Gur, and A. Zeevi. Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems*, 9(4):319–337, 2019.

I. Bogunovic, A. Losalka, A. Krause, and J. Scarlett. Stochastic linear bandits robust to adversarial attacks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 991–999, Online, Apr. 2021.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

X. Chen, H. Zhong, Z. Yang, Z. Wang, and L. Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *Proceedings of the International Conference on Machine Learning*, volume 162, pp. 3773–3793, Baltimore, Maryland, Jul. 2022.

J. Cheng, G. Xiong, X. Dai, Q. Miao, Y. Lv, and F-Y Wang. Rime: Robust preference-based reinforcement learning with noisy preferences. *arXiv preprint arXiv:2402.17257*, 2024.

W. Cheung, D. Simchi-Levi, and R. Zhu. Hedging the drift: Learning to optimize under nonstationarity. *Management Science*, 68(3):1696–1713, 2022.

S. R. Chowdhury, A. Kini, and N. Natarajan. Provably robust DPO: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30:4302–4310, Dec. 2017.

G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, et al. Ultrafeedback: Boosting language models with scaled AI feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Q. Ding, C. J. Hsieh, and J. Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 7111–7123, Online, Mar. 2022.

X. Dong, Z. Wu, Q. Ling, and Z. Tian. Byzantine-robust distributed online learning: Taming adversarial participants in an adversarial environment. *IEEE Transactions on Signal Processing*, 72:235–248, 2024.

K. Ethayarajh, Y. Choi, and S. Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the International Conference on Machine Learning*, volume 162, pp. 5988–6008, Baltimore, MD, Jul. 2022.

Z. Feng, D. Parkes, and H. Xu. The intrinsic robustness of stochastic bandits to strategic manipulation. In *Proceedings of the International Conference on Machine Learning*, volume 119, pp. 3092–3101, Online, Jul. 2020.

T. Fu, M. Sharma, P. Torr, S. B. Cohen, D. Krueger, and F. Barez. Poisonbench: Assessing large language model vulnerability to data poisoning. *arXiv preprint arXiv:2410.08811*, 2025.

P. Gajane, T. Urvoy, and F. Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the International Conference on Machine Learning*, volume 37, pp. 218–227, Lille, France, Jul. 2015.

D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

E. Garcelon, B. Roziere, L. Meunier, J. Tarbouriech, O. Teytaud, A. Lazaric, and M. Pirotta. Adversarial attacks on linear contextual bandits. *Advances in Neural Information Processing Systems*, 33:14362–14373, 2020.

A. Garivier and E. Moulines. On upper-confidence bound policies for switching bandit problems. In *Proceedings of the International Conference on Algorithmic Learning Theory*, Espoo, Finland, Oct. 2011.

A. Gleave, M. Dennis, C. Wild, N. Kant, S. Levine, and S. Russell. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

Z. Guan, K. Ji, D. Bucci Jr, T. Hu, J. Palombo, M. Liston, and Y. Liang. Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4036–4043, New York, NY, Feb. 2020.

A. Gupta, T. Koren, and K. Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Conference on Learning Theory*, volume 99, pp. 1562–1578, Phoenix, AZ, Jun. 2019.

S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Y. Huang and Q. Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *Proceedings of the International Conference on Decision and Game Theory for Security*, volume 10, pp. 217–237, Stockholm, Sweden, Oct. 2019.

K.-S. Jun, L. Li, Y. Ma, and X. Zhu. Adversarial attacks on stochastic bandits. *Advances in Neural Information Processing Systems*, 31:3640–3649, Dec. 2018.

Y. Kang, C. J. Hsieh, and T. C. M. Lee. Robust lipschitz bandits to adversarial corruptions. *Advances in Neural Information Processing Systems*, 36:10897–10908, 2023.

C. W. Kim, S. Verma, M. Tec, and M. Tambe. Preference robustness for DPO with applications to public health. *arXiv preprint arXiv:2509.02709*, 2025.

J. Kos and D. Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

F. Li, L. Lai, and S. Cui. On the adversarial robustness of lasso based feature selection. *IEEE Transactions on Signal Processing*, 69:5555–5567, 2021.

J. Li, B. Zhang, Y. Niu, S. Wu, K. Ding, and J. Wu. Online reward poisoning in reinforcement learning with convergence guarantee. *IEEE Transactions on Information Forensics and Security*, 20:8891–8905, 2025.

L. Li, W. Chu, J. Langford, and R. E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the International Conference on World Wide Web*, pp. 661–670, Raleigh, NC, Apr. 2010.

Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun. Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748*, 2017.

F. Liu and N. Shroff. Data poisoning attacks on stochastic bandits. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 4042–4050, Long Beach, CA, Jun. 2019.

G. Liu and L. Lai. Action-manipulation attacks against stochastic bandits: Attacks and defense. *IEEE Transactions on Signal Processing*, 68:5152–5165, 2020.

G. Liu and L. Lai. Efficient action poisoning attacks on linear contextual bandits. *arXiv preprint arXiv:2112.05367*, 2021a.

G. Liu and L. Lai. Provably efficient black-box action poisoning attacks against reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12400–12410, Dec. 2021b.

G. Liu and L. Lai. Efficient adversarial attacks on online multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 36:24401–24433, Dec. 2023.

T. Lykouris, V. Mirrokni, and R. Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the Annual ACM SIGACT Symposium on Theory of Computing*, pp. 114–122, Los Angeles, CA, Jun. 2018.

M. Lyu, C. Yang, G. Liu, and L. Lai. Adversarial post-action attacks on dueling bandits. In *Proceedings of the Annual Asilomar Conference on Signals, Systems, and Computers*, pp. 118–123, Pacific Grove, CA, Oct. 2025.

E. Ma, S. R. Etesami, and P. Rathi. Reward poisoning on federated reinforcement learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Y. Ma and Z. Zhou. Adversarial attacks on adversarial bandits. *arXiv preprint arXiv:2301.12595*, 2023.

Y. Ma, K.-S. Jun, L. Li, and X. Zhu. Data poisoning attacks in contextual bandits. In *Proceedings of the International Conference on Decision and Game Theory for Security*, pp. 186–204, Seattle, WA, Oct. 2018.

Y. Ma, X. Zhang, W. Sun, and X. Zhu. Policy poisoning in batch reinforcement learning and control. *Advances in Neural Information Processing Systems*, 32:14570–14580, Dec. 2019.

D. Mandal, A. Nika, P. Kamalaruban, A. Singla, and G. Radanović. Corruption robust offline reinforcement learning with human feedback. *arXiv preprint arXiv:2402.06734*, 2024.

A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3932–3939, Vancouver, Canada, Sep. 2017.

J. McMahan, Y. Wu, X. Zhu, and Q. Xie. Optimal attack and defense for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14332–14340, Vancouver, Canada, Feb. 2024.

A. Nika, J. Nöther, D. Mandal, P. Kamalaruban, A. Singla, and G. Radanovic. Policy teaching via data poisoning in learning from human preferences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Mai Khao, Thailand, May. 2025.

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

X. Pan, C. Xiao, W. He, S. Yang, J. Peng, M. Sun, J. Yi, Z. Yang, M. Liu, B. Li, et al. Characterizing attacks on deep reinforcement learning. *arXiv preprint arXiv:1907.09470*, 2019.

P. Pathmanathan, S. Chakraborty, X. Liu, Y. Liang, and F. Huang. Is poisoning a real threat to LLM alignment? maybe more so than you think. *arXiv preprint arXiv:2406.12091*, 2024.

P. Pathmanathan, S. Chakraborty, X. Liu, Y. Liang, and F. Huang. Is poisoning a real threat to DPO? maybe more so than you think. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27556–27564, Philadelphia, PA, Feb. 2025.

A. Pattanaik, Z. Tang, S. Liu, G. Bommannan, and G. Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

K. Pillutla, S. M. Kakade, and Z. Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, Dec. 2023.

A. Rakhsha, G. Radanovic, R. Devidze, X. Zhu, and A. Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *Proceedings of International Conference on Machine Learning*, volume 119, pp. 7974–7984, Online, Jul. 2020.

A. Rakhsha, X. Zhang, X. Zhu, and A. Singla. Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments. *arXiv preprint arXiv:2102.08492*, 2021.

J. Rando and F. Tramèr. Universal jailbreak backdoors from poisoned human feedback. *arXiv preprint arXiv:2311.14455*, 2023.

A. Rangi, H. Xu, L. Tran-Thanh, and M. Franceschetti. Understanding the limits of poisoning attacks in episodic reinforcement learning. *arXiv preprint arXiv:2208.13663*, 2022.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, September 1952.

E. K. Ryu and S. Boyd. Primer on monotone operator methods. *Appl. comput. math*, 15(1):3–43, 2016.

A. Saha, T. Koren, and Y. Mansour. Adversarial dueling bandits. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 9235–9244, Online, Jul. 2021.

A. Saha, A. Pacchiano, and J. Lee. Dueling RL: Reinforcement learning with trajectory preferences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 6263–6289, Valencia, Spain, Apr. 2023.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Y. Sui, M. Zoghi, K. Hofmann, and Y. Yue. Advancements in dueling bandits. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 5502–5510, Stockholm, Sweden, Jul. 2018.

J. Sun, T. Zhang, X. Xie, L. Ma, Y. Zheng, K. Chen, and Y. Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5883–5891, New York, NY, Feb. 2020a.

Y. Sun, D. Huo, and F. Huang. Vulnerability-aware poisoning mechanism for online RL with unknown dynamics. *arXiv preprint arXiv:2009.00774*, 2020b.

C. Tessler, Y. Efroni, and S. Mannor. Action robust reinforcement learning and applications in continuous control. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 6215–6224, Long Beach, CA, Jun. 2019.

Y. Wang, Q. Liu, and C. Jin. Is RLHF more difficult than standard RL? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032, 2023.

J. Wu, J. Wang, C. Xiao, C. Wang, N. Zhang, and Y. Vorobeychik. Preference poisoning attacks on reward model learning. *arXiv preprint arXiv:2402.01920*, 2024a.

J. Wu, Y. Xie, Z. Yang, J. Wu, J. Chen, J. Gao, B. Ding, X. Wang, and X. He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*, 2024b.

J. Wu, J. Wang, C. Xiao, C. Wang, N. Zhang, and Y. Vorobeychik. Preference poisoning attacks on reward model learning. In *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 1622–1640, San Francisco, CA, May. 2025.

R. Wu and W. Sun. Making RL with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*, 2024.

Y. Wu, J. McMahan, X. Zhu, and Q. Xie. Reward poisoning attacks on offline multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10426–10434, Washington, DC, Feb. 2023.

W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

H. Xu, R. Wang, L. Raizman, and Z. Rabinovich. Transferable environment poisoning: Training-time attack on reinforcement learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1398–1406, Online, May. 2021.

Y. Xu, Q. Zeng, and G. Singh. Efficient reward poisoning attacks on online deep reinforcement learning. *arXiv preprint arXiv:2205.14842*, 2022.

Z. Xu, S. Vemuri, K. Panaganti, D. Kalathil, R. Jain, and D. Ramachandran. Robust LLM alignment via distributionally robust direct preference optimization. *arXiv preprint arXiv:2502.01930*, 2025.

Y. Yan, X. Lou, J. Li, Y. Zhang, J. Xie, C. Yu, Y. Wang, D. Yan, and Y. Shen. Reward-robust RLHF in LLMs. *arXiv preprint arXiv:2409.15360*, 2024.

C. Yang, G. Liu, and L. Lai. Stochastic bandits with non-stationary rewards: Reward attack and defense. *IEEE Transactions on Signal Processing*, 72:5007–5020, 2024.

C. Yang, M. Lyu, G. Liu, and L. Lai. Human feedback attack on online RLHF: Attack and robust defense. *IEEE Transactions on Signal Processing*, 73:3886–3901, 2025.

C. Yang, W. Xu, and L. Lai. Efficient preference poisoning attack on offline RLHF. In *Proceedings of the International Conference on Machine Learning*, Seoul, South Korea, Jul. 2026.

L. Yang, M. Hajiesmaili, S. Talebi, J. C. S. Lui, and W. S. Wong. Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm. In *Advances in Neural Information Processing Systems*, volume 33, Dec. 2020.

C. Ye, R. Yang, Q. Gu, and T. Zhang. Corruption-robust offline reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems*, 36:36208–36221, Dec. 2023.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

W. Zhan, M. Uehara, W. Sun, and J. D. Lee. Provable reward-agnostic preference-based reinforcement learning. *arXiv preprint arXiv:2305.18505*, 2023.

X. Zhang, Y. Ma, A. Singla, and X. Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, volume 119, pp. 11225–11234, Online, Jul. 2020.

X. Zhang, Y. Chen, X. Zhu, and W. Sun. Corruption-robust offline reinforcement learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 5757–5773, Online, Mar. 2022.

X. Zhou, Y. Wu, and F. Orabona. A unified theoretical analysis of private and robust offline alignment: from RLHF to DPO. *arXiv preprint arXiv:2505.15694*, 2025.

B. Zhu, M. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 43037–43067, Honolulu, HI, Jul. 2023.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

J. Zimmert and Y. Seldin. An optimal algorithm for stochastic and adversarial bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 467–475, Okinawa, Japan, Apr. 2019.

M. Zoghi, S. Whiteson, R. Munos, and M. Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the International Conference on Machine Learning*, volume 32, pp. 10–18, Beijing, China, Jun. 2014.

A Introduction: Additional Related Work

A.1 Bandit Problems: Adversarial Attacks and Robust Defenses

Adversarial attacks have been studied in several bandit settings, including multi-armed bandits (MAB) (Jun et al., 2018; Liu & Shroff, 2019; Liu & Lai, 2020; Ma & Zhou, 2023; Yang et al., 2024), linear contextual bandits (Ma et al., 2018; Garcelon et al., 2020; Liu & Lai, 2021a), and dueling bandits (Gajane et al., 2015; Agarwal et al., 2021; Saha et al., 2021). These works show that bandit algorithms can be sensitive to adversarially corrupted feedback or interaction signals.

In MAB, the learner chooses one arm in each round and observes the corresponding reward (Robbins, 1952; Auer et al., 1995; 2002). An adversary may corrupt rewards (Jun et al., 2018; Ma & Zhou, 2023; Liu & Shroff, 2019), or manipulate the selected actions (Liu & Lai, 2020) to force the learner to pull a target arm more often. Most existing studies focus on stochastic rewards, where each arm has a fixed reward distribution. Other works consider adversarial rewards, where rewards are chosen arbitrarily (Auer et al., 2002; Zimmert & Seldin, 2019), or non-stationary rewards, where reward distributions change over time within some extent (Yang et al., 2024; Garivier & Moulines, 2011; Besbes et al., 2014; Auer et al., 2019; Besbes et al., 2019; Cheung et al., 2022).

In linear contextual bandits, the expected reward is assumed to be linear in the context features (Auer, 2002; Abe et al., 2003; Li et al., 2010). Existing attacks include reward poisoning (Ma et al., 2018; Garcelon et al., 2020), context poisoning (Garcelon et al., 2020), and action poisoning (Liu & Lai, 2021a).

Dueling bandits are closer to RLHF, in which the learner selects a pair of arms and observes only which one is preferred (Yue et al., 2012; Zoghi et al., 2014; Sui et al., 2018). Prior work studies stochastic dueling bandits with corrupted pairs (Agarwal et al., 2021), as well as adversarial preference models (Gajane et al., 2015; Saha et al., 2021; Lyu et al., 2025).

Adversarial attacks naturally motivate robust bandit algorithms. Existing robust defenses have been studied for stochastic bandits (Lykouris et al., 2018; Gupta et al., 2019; Guan et al., 2020; Bogunovic et al., 2021), adversarial bandits (Yang et al., 2020), stochastic linear contextual bandits (Ding et al., 2022), strategic manipulation in stochastic bandits (Feng et al., 2020), stochastic Lipschitz bandits (Kang et al., 2023), and non-stationary bandits (Yang et al., 2024). These defenses are important in real applications, where corrupted feedback can lead to harmful decisions such as unfair recommendations or unsafe medical treatments.

A.2 Standard Reinforcement Learning: Adversarial Attacks and Robust Defenses

Adversarial attacks on standard RL have also been widely studied, in which the attacker corrupts multiple aspects of the learning process so that the learned policy is misled. Common poisoning models include reward/cost poisoning (Huang & Zhu, 2019; Zhang et al., 2020; Rakhsha et al., 2020; 2021; Banihashem et al., 2022; Xu et al., 2022; Li et al., 2025); action poisoning (Liu & Lai, 2021b); and more general environment or transition dynamics poisoning (Rakhsha et al., 2020; Xu et al., 2021). Poisoning attacks are also studied in batch settings (Ma et al., 2019), and in dynamics-agnostic settings (Sun et al., 2020b; Liu & Lai, 2021b; Rakhsha et al., 2021; Xu et al., 2022; Li et al., 2025).

Beyond single-agent RL, adversarial attacks have also been considered in multi-agent RL (Liu & Lai, 2023; Wu et al., 2023), and in federated RL (Ma et al., 2024). Other works characterize the limits of reward-only or action-only poisoning in episodic RL (Rangi et al., 2022), or formulate the optimal attack through the attacker's own Markov decision process (McMahan et al., 2024). In deep RL, the policies are neural networks and the attackers modify the high-dimensional observations (Huang et al., 2017; Kos & Song, 2017; Lin et al., 2017; Pattanaik et al., 2017; Mandlekar et al., 2017; Pan et al., 2019; Gleave et al., 2019; Sun et al., 2020a).

Existing works have proposed defense strategies in RL algorithms (Tessler et al., 2019; Zhang et al., 2022; Ye et al., 2023). Related robustness ideas have also been developed in broader machine learning settings (Dong et al., 2024; Li et al., 2021; Pillutla et al., 2022). These works show that defense mechanisms are necessary when corrupted interactions can push the learner toward undesirable behaviors.

A.3 Reinforcement Learning from Human Feedback: Adversarial Attacks and Robust Defenses

RLHF is different from bandit and standard RL because it learns policies from pairwise discrete preference data rather than a numerical reward from the environment. Depending on whether the training is on a pre-collected dataset or through real-time interaction with human annotators (Xiong et al., 2023), RLHF can be categorized into offline RLHF (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Zhu et al., 2023; Rafailov et al., 2023) and online RLHF (Chen et al., 2022; Saha et al., 2023; Wang et al., 2023; Zhan et al., 2023; Wu & Sun, 2024). It has been shown that offline RLHF is vulnerable to preference poisoning, such as data injection attack and label flipping attack, through experiments (Rando & Tramèr, 2023; Wu et al., 2024a; Baumgärtner et al., 2024; Pathmanathan et al., 2024; Fu et al., 2025; Pathmanathan et al., 2025; Wu et al., 2025), and theoretical analysis (Nika et al., 2025). Nika et al. (2025) proves that a large poisoned preference set may be needed to mislead offline RLHF for data injection attack. Meanwhile, Yang et al. (2026) proposes efficient label flipping attacks that strategically select a small subset of preference pairs to flip and mislead offline RLHF toward a target policy. In online RLHF, preferences can also be attacked to manipulate the learning process (Yang et al., 2025).

Other than the adversarial attack described above, RLHF is also sensitive to the random noise in preference data, which is caused by annotation errors, limited expert involvement, and other non-adversarial sources (Casper et al., 2023; Yan et al., 2024). This noise may lead to unpredictable or biased aligned model behavior.

Motivated by these reliability issues of preference data, existing works propose robust online RLHF against adversarial attacks (Yang et al., 2025) and robust offline RLHF against random noise (Yan et al., 2024; Cheng et al., 2024; Mandal et al., 2024; Chowdhury et al., 2024; Wu et al., 2024b) and distribution shift (Xu et al., 2025). In the offline setting, there is a gap in the literature on robust RLHF against adversarial attacks, especially for specifically designed attacks leading to a target policy. This paper aims to fill this gap by proposing robust DPO methods against budgeted preference label flips, including adversarial flips.

B Problem Formulation: Why the Bounded Training Region Matters

Remark B.1 (Why the bounded-region Assumption 2.3 matters). Without Assumption 2.3, the standard DPO objective need not be globally strongly convex on all of $\mathbb{R}^d$. In the log-linear setting, the per-sample Hessian is

$$\nabla^2 \ell_i(\theta) = \beta^2 \, \sigma(z_i(\theta))\big(1 - \sigma(z_i(\theta))\big)\, v_i v_i^\top,$$

and $\inf_\theta \sigma(z_i(\theta))\big(1 - \sigma(z_i(\theta))\big) = 0$ as $\|\theta - \theta_{\mathrm{ref}}\|_2 \to \infty$ along separable directions. Thus, global strong convexity is not guaranteed even though the bounded-region curvature constant μ_G in (12) is positive.

The strong-convexity statement in Nika et al. (2025) does not contradict the above discussion. The key distinction is whether the curvature comes from the logistic part on a bounded training region or from an explicit quadratic regularizer. In the notation of Nika et al. (2025), their parameter displacement is

$$v := \theta - \theta_\mu,$$

and their regularized DPO/logistic objective includes the quadratic term

$$\frac{\lambda}{2}\|\theta - \theta_\mu\|_2^2.$$

For the unregularized logistic part, their Hessian lower bound contains a data covariance term multiplied by a logistic-curvature constant:

$$\nabla^2 \ell_v(D) \succeq n\beta C_v \Sigma_D,$$

where

$$C_v := \min_i \frac{\exp(\beta y_i v^\top x_i + b)}{\big(1 + \exp(\beta y_i v^\top x_i + b)\big)^2}.$$

Here, their C_v is exactly the minimum logistic curvature factor over the dataset. In our notation, their center θ_μ corresponds to the reference parameter θ_{ref}, their displacement v corresponds to $\theta - \theta_{\mathrm{ref}}$, and their signed feature $y_i x_i$ corresponds to our pairwise feature direction

$$v_i = \psi(x^i, y_w^i) - \psi(x^i, y_l^i),$$

up to the sign convention used to write the logistic loss. Thus their scalar argument $\beta y_i v^\top x_i + b$ corresponds to our DPO logit $z_i(\theta)$.

At any fixed finite parameter displacement, $C_v > 0$. More importantly for our setting, if training is restricted to a bounded ball around the reference parameter, then compactness of this ball and finiteness of the dataset imply a positive lower bound

$$\underline{C}_R := \inf_{\|\theta - \theta_{\mathrm{ref}}\|_2 \le R} \min_i \frac{\exp(z_i(\theta))}{(1 + \exp(z_i(\theta)))^2} > 0. \tag{18}$$

Consequently, on the bounded training region, the unregularized DPO objective is strongly convex whenever the corresponding empirical feature covariance is positive definite on the trained parameter subspace. This is the local curvature used in our unregularized Trim-K and Soft-Trim-K bounds.

This local statement does not imply global strong convexity on all of $\mathbb{R}^d$. Indeed, for

$$q(z) := \frac{e^z}{(1 + e^z)^2},$$

we have $q(z) > 0$ for every finite z, but $q(z) \to 0$ as $z \to \pm\infty$. Therefore, along a direction where some sample logit diverges, the corresponding logistic curvature factor vanishes, and the global infimum of C_v is zero. This is why the unregularized objective need not admit a uniform global strong-convexity constant.

The quadratic regularizer in Nika et al. (2025) gives a different source of curvature. It adds λI to the Hessian, so the regularized objective can obtain an additional global curvature term even when the logistic curvature factor is small. This is the same mechanism used by our regularized variants in Section 3.4: the denominator improves by λ, while comparison to the unregularized clean DPO target also introduces the deterministic regularization bias term λR.

Proposition B.2 (Linear-in-T deviation under convex smoothness alone)**.** *For any $\bar{L} \in (0, \beta^2]$, there exist a convex $\bar{L}$-smooth function F, a constant bias vector $b \neq 0$, and a step size $0 < \eta \le 2/\bar{L}$ such that, when clean and attacked gradient descent are run from the same initialization,*

$$\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}} - \eta \nabla F(\theta_t^{\mathrm{cln}}), \qquad \theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta(\nabla F(\theta_t^{\mathrm{atk}}) + b),$$

the iterate deviation satisfies

$$\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 = \eta T \|b\|_2 \qquad \textit{for all } T \ge 1.$$

Thus, convex smoothness plus a uniform gradient-bias bound is not enough to imply a time-uniform deviation bound.

Proof of Proposition B.2. Choose any nonzero vector b and any $0 < \eta \le 2/\bar{L}$. Take $F(\theta) \equiv 0$ for all θ. Then F is convex and $\bar{L}$-smooth for every $\bar{L} > 0$, and $\nabla F(\theta) = 0$ for all θ. The clean GD iterates are constant:

$$\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}}.$$

The attacked GD iterates satisfy

$$\theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta b.$$

Starting from the same initialization $\theta_0^{\mathrm{atk}} = \theta_0^{\mathrm{cln}}$, we obtain by induction

$$\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}} = -\eta T b,$$

and hence $\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 = \eta T \|b\|_2$. □

Proposition B.2 explains why Assumption 2.3 is important for time-uniform robustness. Without a bounded training region, the DPO objective may not have a uniform curvature lower bound, so the stability argument would fall back to the accumulated-bias recursion over gradient steps, leading to a deviation that grows linearly in T, illustrated by the proposition. The bounded-region assumption prevents this failure mode by keeping the logistic curvature uniformly positive along training. For the regularized variants, the quadratic regularizer provides an additional curvature term, but its additional bias means that it need not improve the complete deviation bound.

C METHODS: ALGORITHM DETAILS

Algorithms 1 and 2 give the implementation details for applying the objective-level defenses described in the main text to the observed dataset. In the log-linear setting, the score $s_i = \beta\|v_i\|_2$ is the exact magnitude of a pair's potential flip-induced gradient change. It does not estimate whether the pair is corrupted. Because $\|\psi(x, y_w) - \psi(x, y_l)\|_2$ is invariant to swapping the two responses, this score can be computed directly from the observed pair even if it has been flipped. Given precomputed d-dimensional features, computing all N scores costs $O(Nd)$ and sorting them costs $O(N \log N)$, once before training. Training then uses standard DPO updates on the fixed retained set or with fixed scalar weights on the per-example losses. The REG variants additionally evaluate the quadratic penalty and its gradient, requiring $O(d)$ arithmetic per update.

Algorithm 1 Trim-K Robust-DPO

Require: Observed preference dataset $\tilde{\mathcal{D}}(A) = \{(x^i, \tilde{y}_w^i, \tilde{y}_l^i)\}_{i=1}^N$, budget upper bound K, reference policy π_{ref}, inverse-temperature β, optimizer, training horizon T

for $i = 1, \ldots, N$ **do**
  Compute observed direction $\tilde{v}_i \leftarrow \psi(x^i, \tilde{y}_w^i) - \psi(x^i, \tilde{y}_l^i)$
  Compute score $s_i \leftarrow \beta\|\tilde{v}_i\|_2$
end for
Sort $\{s_i\}_{i=1}^N$ in descending order and keep the $N - K$ smallest-score indices G
Initialize parameter θ_0
for $t = 0, 1, \ldots, T-1$ **do**
  Compute the observed trimmed DPO objective

$$\tilde{F}_{G,A}(\theta_t) = \frac{1}{N-K} \sum_{i \in G} \ell_{\text{DPO}}(\theta_t; x^i, \tilde{y}_w^i, \tilde{y}_l^i)$$

  Update $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \tilde{F}_{G,A}(\theta_t)$
end for
return θ_T

Algorithm 2 Soft-Trim-K Robust-DPO

Require: Observed preference dataset $\tilde{\mathcal{D}}(A) = \{(x^i, \tilde{y}_w^i, \tilde{y}_l^i)\}_{i=1}^N$, budget upper bound K, temperature τ, reference policy π_{ref}, inverse-temperature β, optimizer, training horizon T

for $i = 1, \ldots, N$ **do**
  Compute observed direction $\tilde{v}_i \leftarrow \psi(x^i, \tilde{y}_w^i) - \psi(x^i, \tilde{y}_l^i)$
  Compute score $s_i \leftarrow \beta\|\tilde{v}_i\|_2$
end for
Rank the scores in descending order and assign weights $w_i^{(\tau)}$ by (13)
Initialize parameter θ_0
for $t = 0, 1, \ldots, T-1$ **do**
  Compute the observed weighted DPO objective

$$\tilde{F}_{\tau,A}(\theta_t) = \frac{1}{W_\tau} \sum_{i=1}^{N} w_i^{(\tau)} \ell_{\text{DPO}}(\theta_t; x^i, \tilde{y}_w^i, \tilde{y}_l^i)$$

  Update $\theta_{t+1} \leftarrow \theta_t - \eta \nabla \tilde{F}_{\tau,A}(\theta_t)$
end for
return θ_T

D METHODS: SOFT-TRIM WEIGHT VISUALIZATION

Figure 2 visualizes the rank-based weights used by Soft-Trim-K Robust-DPO. Smaller values of τ more aggressively downweight the highest-score samples and approach Trim-K Robust-DPO, while larger values of τ keep the weights closer to one and approach standard DPO.

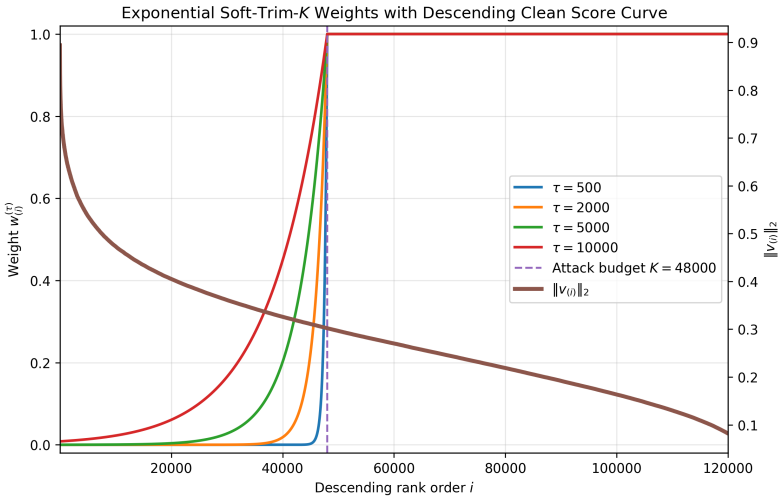


Figure 2: Left vertical axis: Soft-Trim-K weights $w_i^{(\tau)}$ as a function of the rank $r(i)$, for different temperatures τ. Right vertical axis: the impact scores $s_i = \beta\|v_i\|_2$ as a function of rank, with $\beta = 1$. The trimming cutoff at $K = 48{,}000$ is indicated by the vertical dashed line.

E METHODS: SUPPORTING LEMMAS FOR THE MAIN ANALYSIS

This appendix collects the standard smoothness, curvature, and optimization facts used in Section 3.

Lemma E.1 (Smoothness of per-sample and normalized log-linear DPO objectives)**.** *Under Assumption 2.1, for every sample i, the clean per-sample loss ℓ_i defined in* (2) *is convex and L_i-smooth, where*

$$L_i \le \frac{\beta^2}{4}\|v_i\|_2^2 \le \beta^2.$$

In particular, for all θ, θ',

$$\|\nabla\ell_i(\theta) - \nabla\ell_i(\theta')\|_2 \le \beta^2\|\theta - \theta'\|_2.$$

The same conclusion holds for the observed loss $\tilde{\ell}_i$, because swapping the preference orientation replaces v_i by $-v_i$ and preserves its norm. Consequently, for any fixed weights $a_i \ge 0$ with $\sum_i a_i = 1$, both normalized objectives

$$F_a(\theta) := \sum_i a_i\ell_i(\theta), \qquad \tilde{F}_a(\theta) := \sum_i a_i\tilde{\ell}_i(\theta)$$

are convex and β^2-smooth. In particular, this applies to F, F_G, $\tilde{F}_{G,A}$, F_τ, and $\tilde{F}_{\tau,A}$.

Lemma E.2 (Restricted strong convexity of the trimmed DPO objective)**.** *Under Assumptions 2.1 and 2.3, the clean trimmed objective F_G introduced in Section 3.2 is μ_G-strongly convex on $\mathbb{B}_R(\theta_{\mathrm{ref}})$ and β^2-smooth, where μ_G is defined in* (12)*.*

Lemma E.3 (Closed-form bound on the clean Trim-K bias)**.** *Under Assumption 2.1, for every $\theta \in \mathbb{R}^d$,*

$$\|\nabla F_G(\theta) - \nabla F(\theta)\|_2 \;\le\; \frac{1}{N}\left(K\,\delta(K) + \sum_{j=1}^{K} s_{(j)}\right).$$

Consequently,

$$\mathsf{B}_G^{\mathrm{cln}}(R) \;\le\; \frac{K\,\delta(K)+\sum_{j=1}^{K} s_{(j)}}{N} \;\le\; \frac{2}{N}\sum_{j=1}^{K} s_{(j)} \;\le\; \frac{4\beta K}{N}.$$

Moreover, for the regularized clean bias $\mathsf{B}_{G,\lambda}^{\mathrm{cln}}(R)$*,*

$$\mathsf{B}_{G,\lambda}^{\mathrm{cln}}(R) \;\le\; \frac{K\,\delta(K)+\sum_{j=1}^{K} s_{(j)}}{N} + \lambda R.$$

Lemma E.4 (Closed-form bound on the clean Soft-Trim-K bias)**.** *Define*

$$\mathsf{C}_\tau := \frac{1}{NW_\tau}\left[\sum_{j=1}^{K}\left|D_\tau - N\big(1-a_j^{(\tau)}\big)\right| s_{(j)} + D_\tau \sum_{j=K+1}^{N} s_{(j)}\right]. \tag{19}$$

Under Assumption 2.1, for every $\theta \in \mathbb{R}^d$*,*

$$\|\nabla F_\tau(\theta) - \nabla F(\theta)\|_2 \;\le\; \mathsf{C}_\tau.$$

Consequently,

$$\mathsf{B}_\tau^{\mathrm{cln}}(R) \;\le\; \mathsf{C}_\tau \;\le\; \frac{D_\tau \bar{s} + \sum_{j=1}^{K}(1-a_j^{(\tau)})s_{(j)}}{W_\tau} \;\le\; \frac{4\beta K}{W_\tau}.$$

Moreover, for the regularized clean bias $\mathsf{B}_{\tau,\lambda}^{\mathrm{cln}}(R)$*,*

$$\mathsf{B}_{\tau,\lambda}^{\mathrm{cln}}(R) \;\le\; \mathsf{C}_\tau + \lambda R.$$

As $\tau \downarrow 0$*,* C_τ *reduces to the exact hard-trim type bound*

$$\frac{1}{N(N-K)}\left[(N-K)\sum_{j=1}^{K} s_{(j)} + K\sum_{j=K+1}^{N} s_{(j)}\right],$$

which is at most $\big(K\delta(K)+\sum_{j=1}^{K} s_{(j)}\big)/N$*. As* $\tau \to \infty$*, all weights converge to one,* $D_\tau \to 0$*, and* $\mathsf{C}_\tau \to 0$*.*

Lemma E.5 (Non-expansiveness of gradient descent for convex smooth objectives)**.** *Let* $F : \mathbb{R}^d \to \mathbb{R}$ *be convex and* L*-smooth. Define* $T(\theta) := \theta - \eta\nabla F(\theta)$*. If* $0 < \eta \le 2/L$*, then* T *is non-expansive:*

$$\|T(x) - T(y)\|_2 \le \|x - y\|_2, \qquad \forall x, y \in \mathbb{R}^d.$$

Lemma E.6 (Time-uniform stability under objective perturbation)**.** *Let* $F_0 : \mathbb{R}^d \to \mathbb{R}$ *be the target objective and let* $H : \mathbb{R}^d \to \mathbb{R}$ *be the objective optimized by the attacked run. Assume* H *is* μ*-strongly convex and* L*-smooth on a convex region containing the clean and attacked iterates, and assume*

$$\sup_{\theta \text{ in the region}} \|\nabla H(\theta) - \nabla F_0(\theta)\|_2 \le B.$$

Run full-batch gradient descent from the same initialization with step size $0 < \eta \le 1/L$*:*

$$\theta_{t+1}^{\mathrm{cln}} = \theta_t^{\mathrm{cln}} - \eta\nabla F_0(\theta_t^{\mathrm{cln}}), \qquad \theta_{t+1}^{\mathrm{atk}} = \theta_t^{\mathrm{atk}} - \eta\nabla H(\theta_t^{\mathrm{atk}}).$$

Then for every $T \ge 1$*,*

$$\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2 \le \frac{B}{\mu}\big(1-(1-\eta\mu)^T\big) \le \frac{B}{\mu}.$$

F Methods: Proofs of Theoretical Results

F.1 Proof of Lemma E.1

Proof. Fix a sample i. Since b_i does not depend on θ, $\nabla_\theta z_i(\theta) = \beta v_i$. Differentiating,

$$\nabla\ell_i(\theta) = -\big(1-\sigma(z_i(\theta))\big)\,\beta v_i, \qquad \nabla^2\ell_i(\theta) = \beta^2\sigma(z_i(\theta))\big(1-\sigma(z_i(\theta))\big)\,v_i v_i^\top.$$

Thus, using $\sigma(z_i(\theta))(1-\sigma(z_i(\theta))) \le 1/4$ and $\|v_i v_i^\top\|_{\mathrm{op}} = \|v_i\|_2^2$,

$$\|\nabla^2 \ell_i(\theta)\|_{\mathrm{op}} \le \frac{\beta^2}{4}\|v_i\|_2^2.$$

Under Assumption 2.1, $\|v_i\|_2 \le \|\psi(x^i, y_w^i)\|_2 + \|\psi(x^i, y_l^i)\|_2 \le 2$, hence $\|\nabla^2 \ell_i(\theta)\|_{\mathrm{op}} \le \beta^2$ for all θ. A uniform Hessian bound implies gradient Lipschitzness:

$$\|\nabla \ell_i(\theta) - \nabla \ell_i(\theta')\|_2 \le \sup_{\xi} \|\nabla^2 \ell_i(\xi)\|_{\mathrm{op}} \, \|\theta - \theta'\|_2 \le \beta^2 \|\theta - \theta'\|_2.$$

The Hessian is positive semidefinite, so ℓ_i is convex. For the observed orientation, either the pair is unchanged or (v_i, b_i, z_i) becomes $(-v_i, -b_i, -z_i)$. The same Hessian bound and convexity conclusion therefore hold for $\tilde{\ell}_i$.

Finally, for any fixed normalized nonnegative weighted average $F_a(\theta) = \sum_i a_i \ell_i(\theta)$,

$$\|\nabla^2 F_a(\theta)\|_{\mathrm{op}} = \Big\| \sum_i a_i \nabla^2 \ell_i(\theta) \Big\|_{\mathrm{op}} \le \sum_i a_i \|\nabla^2 \ell_i(\theta)\|_{\mathrm{op}} \le \beta^2,$$

and $\nabla^2 F_a(\theta) \succeq 0$. The identical argument applies to $\tilde{F}_a$ using the observed losses $\tilde{\ell}_i$. This proves the stated claims for the fixed normalized clean and attacked empirical objectives. □

F.2 Proof of Lemma E.2

Proof. For $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$, the Hessian of the clean trimmed objective is

$$\nabla^2 F_G(\theta) = \frac{1}{|G|} \sum_{i \in G} \beta^2 \sigma(z_i(\theta)) \big(1 - \sigma(z_i(\theta))\big) v_i v_i^\top.$$

By the definition of $\underline{c}_G(R)$,

$$\nabla^2 F_G(\theta) \succeq \beta^2 \underline{c}_G(R) \frac{1}{|G|} \sum_{i \in G} v_i v_i^\top = \beta^2 \underline{c}_G(R) \Sigma_G \succeq \mu_G I.$$

Therefore F_G is μ_G-strongly convex on the bounded region. Its β^2-smoothness follows from Lemma E.1. □

F.3 Proof of Lemma E.3

Proof. Let $T := [N] \setminus G$ denote the trimmed set, so $|T| = K$ and $|G| = N - K$. Splitting $\sum_{i=1}^N g_i(\theta) = \sum_{i \in G} g_i(\theta) + \sum_{i \in T} g_i(\theta)$, we obtain

$$\begin{aligned}
\nabla F_G(\theta) - \nabla F(\theta) &= \frac{1}{N-K} \sum_{i \in G} g_i(\theta) - \frac{1}{N} \sum_{i=1}^N g_i(\theta) \\
&= \left(\frac{1}{N-K} - \frac{1}{N} \right) \sum_{i \in G} g_i(\theta) - \frac{1}{N} \sum_{i \in T} g_i(\theta) \\
&= \frac{K}{N(N-K)} \sum_{i \in G} g_i(\theta) - \frac{1}{N} \sum_{i \in T} g_i(\theta).
\end{aligned}$$

By the closed-form gradient (4) and $|\sigma(z) - 1| \le 1$, we have $\|g_i(\theta)\|_2 \le \beta \|v_i\|_2 = s_i$ for every θ. Triangle inequality gives

$$\|\nabla F_G(\theta) - \nabla F(\theta)\|_2 \le \frac{K}{N(N-K)} \sum_{i \in G} s_i + \frac{1}{N} \sum_{i \in T} s_i.$$

By construction of G, every $i \in G$ satisfies $s_i \le \delta(K)$, hence $\sum_{i \in G} s_i \le (N-K)\delta(K)$. Also, T contains exactly the K largest scores, so $\sum_{i \in T} s_i = \sum_{j=1}^K s_{(j)}$. Substituting,

$$\|\nabla F_G(\theta) - \nabla F(\theta)\|_2 \le \frac{K\,\delta(K)}{N} + \frac{1}{N} \sum_{j=1}^K s_{(j)} = \frac{K\,\delta(K) + \sum_{j=1}^K s_{(j)}}{N}.$$

Taking the supremum over $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$ gives the bound on $\mathsf{B}_G^{\mathrm{cln}}(R)$. Since $\delta(K) = s_{(K+1)} \le s_{(j)}$ for $j \le K$, $K\delta(K) \le \sum_{j=1}^K s_{(j)}$, yielding the simplified $\frac{2}{N}\sum_{j=1}^K s_{(j)}$. Each $s_i = \beta\|v_i\|_2 \le 2\beta$ under Assumption 2.1, hence $\sum_{j=1}^K s_{(j)} \le 2\beta K$, giving the universal bound $\frac{4\beta K}{N}$. For the regularized bias, note

$$\nabla F_G(\theta) + \lambda(\theta - \theta_{\mathrm{ref}}) - \nabla F(\theta) = (\nabla F_G(\theta) - \nabla F(\theta)) + \lambda(\theta - \theta_{\mathrm{ref}}),$$

and $\|\theta - \theta_{\mathrm{ref}}\|_2 \le R$ on $\mathbb{B}_R(\theta_{\mathrm{ref}})$, so the regularized bound follows by the triangle inequality. □

F.4 Proof of Lemma E.4

Proof. Write the samples in descending score order, so that the top-K weights are $a_j^{(\tau)}$ for $j \le K$ and the remaining weights are one. Recall from Section 3.3 that $D_\tau = K - \sum_{j=1}^K a_j^{(\tau)} = N - W_\tau$. Since

$$\nabla F_\tau(\theta) - \nabla F(\theta) = \sum_{i=1}^N \left(\frac{w_i^{(\tau)}}{W_\tau} - \frac{1}{N}\right) g_i(\theta),$$

and $\|g_i(\theta)\|_2 \le s_i$, the triangle inequality gives

$$\|\nabla F_\tau(\theta) - \nabla F(\theta)\|_2 \le \sum_{i=1}^N \left|\frac{w_i^{(\tau)}}{W_\tau} - \frac{1}{N}\right| s_i.$$

For a top-K sample with rank j, the absolute coefficient is

$$\left|\frac{a_j^{(\tau)}}{W_\tau} - \frac{1}{N}\right| = \frac{\left|Na_j^{(\tau)} - W_\tau\right|}{NW_\tau} = \frac{\left|D_\tau - N(1 - a_j^{(\tau)})\right|}{NW_\tau}.$$

For a sample with rank $j > K$, the coefficient is

$$\left|\frac{1}{W_\tau} - \frac{1}{N}\right| = \frac{D_\tau}{NW_\tau}.$$

Combining these two displays proves

$$\|\nabla F_\tau(\theta) - \nabla F(\theta)\|_2 \le \mathsf{C}_\tau.$$

Taking the supremum over $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$ gives $\mathsf{B}_\tau^{\mathrm{cln}}(R) \le \mathsf{C}_\tau$. Also, for every $j \le K$,

$$\left|D_\tau - N(1 - a_j^{(\tau)})\right| \le D_\tau + N(1 - a_j^{(\tau)}),$$

which implies

$$\mathsf{C}_\tau \le \frac{D_\tau \bar{s} + \sum_{j=1}^K (1 - a_j^{(\tau)}) s_{(j)}}{W_\tau}.$$

The same simpler upper bound also follows directly from the alternative decomposition

$$\nabla F_\tau(\theta) - \nabla F(\theta) = \frac{1}{W_\tau}\left[D_\tau \nabla F(\theta) - \sum_{j=1}^K (1 - a_j^{(\tau)}) g_{(j)}(\theta)\right],$$

where $g_{(j)}$ denotes the gradient of the sample with the j-th largest score. Using $\|\nabla F(\theta)\|_2 \le \bar{s}$ and $\|g_{(j)}(\theta)\|_2 \le s_{(j)}$ gives

$$\mathsf{B}_\tau^{\mathrm{cln}}(R) \le \frac{D_\tau \bar{s} + \sum_{j=1}^K (1 - a_j^{(\tau)}) s_{(j)}}{W_\tau}.$$

Finally, $\bar{s} \le 2\beta$, $D_\tau = \sum_{j=1}^K (1 - a_j^{(\tau)}) \le K$, and $s_{(j)} \le 2\beta$ imply the universal bound $4\beta K / W_\tau$. The regularized bound follows by adding $\lambda(\theta - \theta_{\mathrm{ref}})$ and using $\|\theta - \theta_{\mathrm{ref}}\|_2 \le R$. The two limiting statements follow by taking $a_j^{(\tau)} \to 0$ as $\tau \downarrow 0$ and $a_j^{(\tau)} \to 1$ as $\tau \to \infty$ in the definition of C_τ. □

F.5 Proof of Lemma E.5

Proof. Let $d := x - y$ and $\Delta g := \nabla F(x) - \nabla F(y)$. We first recall that for convex L-smooth F, ∇F is $1/L$-cocoercive (Bauschke & Combettes, 2009):

$$\langle d, \Delta g \rangle \;\geq\; \frac{1}{L}\|\Delta g\|_2^2. \tag{20}$$

Now expand the distance after one GD step:

$$\|T(x) - T(y)\|_2^2 = \|d - \eta \Delta g\|_2^2 = \|d\|_2^2 - 2\eta\langle d, \Delta g\rangle + \eta^2\|\Delta g\|_2^2.$$

Applying (20) yields

$$\|T(x) - T(y)\|_2^2 \leq \|d\|_2^2 + \Big(\eta^2 - \frac{2\eta}{L}\Big)\|\Delta g\|_2^2.$$

If $0 < \eta \leq 2/L$, then $\eta^2 - \frac{2\eta}{L} \leq 0$, hence $\|T(x) - T(y)\|_2^2 \leq \|d\|_2^2$, which implies $\|T(x) - T(y)\|_2 \leq \|x - y\|_2$. □

F.6 Proof of Lemma E.6

Proof. Let $T_H(\theta) := \theta - \eta \nabla H(\theta)$ and $T_0(\theta) := \theta - \eta \nabla F_0(\theta)$. Since H is μ-strongly convex and L-smooth, and $0 < \eta \leq 1/L$, the gradient-descent map T_H is a contraction (Ryu & Boyd, 2016):

$$\|T_H(x) - T_H(y)\|_2 \leq (1 - \eta\mu)\|x - y\|_2.$$

Therefore, with $\Theta_t := \theta_t^{\mathrm{atk}} - \theta_t^{\mathrm{cln}}$,

$$\begin{aligned}
\|\Theta_{t+1}\|_2 &= \|T_H(\theta_t^{\mathrm{atk}}) - T_0(\theta_t^{\mathrm{cln}})\|_2 \\
&\leq \|T_H(\theta_t^{\mathrm{atk}}) - T_H(\theta_t^{\mathrm{cln}})\|_2 + \eta\|\nabla H(\theta_t^{\mathrm{cln}}) - \nabla F_0(\theta_t^{\mathrm{cln}})\|_2 \\
&\leq (1 - \eta\mu)\|\Theta_t\|_2 + \eta B.
\end{aligned}$$

Unrolling this recursion from $\Theta_0 = 0$ gives

$$\|\Theta_T\|_2 \leq \eta B \sum_{t=0}^{T-1}(1 - \eta\mu)^{T-1-t} = \frac{B}{\mu}\big(1 - (1 - \eta\mu)^T\big) \leq \frac{B}{\mu}.$$

□

F.7 Proof of Lemma 2.2

We include the proof in our notation, following the gradient-shift analysis in Yang et al. (2026).

Proof. For sample i, swapping (y_w^i, y_l^i) changes (v_i, b_i) to $(-v_i, -b_i)$ and changes the logit to $z_i^{\mathrm{flip}}(\theta) = -z_i(\theta)$. Using $\sigma(-z) = 1 - \sigma(z)$ and $g_i(\theta) = \beta(\sigma(z_i(\theta)) - 1)v_i$, we obtain

$$g_i^{\mathrm{flip}}(\theta) = \beta(\sigma(-z_i(\theta)) - 1)(-v_i) = \beta\sigma(z_i(\theta))v_i.$$

Therefore $g_i^{\mathrm{flip}}(\theta) - g_i(\theta) = \beta v_i$. □

F.8 Proof of Theorem 3.1

Proof. By the decomposition before the theorem,

$$\nabla\tilde{F}_{G,A}(\theta) - \nabla F(\theta) = \big(\nabla F_G(\theta) - \nabla F(\theta)\big) + \frac{1}{|G|}\sum_{i \in A \cap G}\beta v_i.$$

The first term is bounded by $\mathsf{B}_G^{\mathrm{cln}}(R)$ by definition. For the second term,

$$\left\|\frac{1}{|G|}\sum_{i \in A\cap G}\beta v_i\right\|_2 \leq \frac{1}{|G|}\sum_{i \in A\cap G}\|\beta v_i\|_2 = \frac{1}{|G|}\sum_{i \in A\cap G} s_i \leq \frac{|A \cap G|}{|G|}\max_{i \in G} s_i \leq \frac{K}{N - K}\delta(K).$$

Combining the two bounds proves the claim. □

F.9 Proof of Theorem 3.2

Proof. The Hessian of $\tilde{F}_{G,A}$ is the same as the Hessian of F_G, because label swaps replace (v_i, z_i) by $(-v_i, -z_i)$ and leave $\sigma(z_i)(1-\sigma(z_i))v_i v_i^\top$ unchanged. Thus, by Lemma E.2, $\tilde{F}_{G,A}$ is μ_G-strongly convex and β^2-smooth on the bounded region. Theorem 3.1 gives the uniform perturbation bound

$$\sup_{\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})} \|\nabla \tilde{F}_{G,A}(\theta) - \nabla F(\theta)\|_2 \le \mathsf{B}_G^{\mathrm{cln}}(R) + \frac{K}{N-K}\delta(K).$$

Applying Lemma E.6 with $F_0 = F$, $H = \tilde{F}_{G,A}$, $\mu = \mu_G$, $L = \beta^2$, and

$$B = \mathsf{B}_G^{\mathrm{cln}}(R) + \frac{K}{N-K}\delta(K)$$

proves the stated bound. □

F.10 Proof of Corollary 3.3

Proof. Fix x and define the log-partition

$$A_x(\theta) := \log \sum_y \exp\big(\langle \theta, \psi(x,y)\rangle\big).$$

Under Assumption 2.1 ($\|\psi(x,y)\|_2 \le 1$), we have $\|\nabla^2 A_x(\theta)\|_{\mathrm{op}} \le 1$ for all θ, hence A_x is 1-smooth. Moreover, $\pi_\theta(\cdot|x)$ is the exponential family induced by A_x, and the conditional KL admits the standard Bregman form

$$D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x) \,\|\, \pi_{\theta'}(\cdot|x)\big) = A_x(\theta') - A_x(\theta) - \langle \nabla A_x(\theta), \theta' - \theta\rangle.$$

By 1-smoothness of A_x, the descent inequality implies that for all θ, θ',

$$A_x(\theta') \le A_x(\theta) + \langle \nabla A_x(\theta),\, \theta' - \theta\rangle + \frac{1}{2}\|\theta' - \theta\|_2^2.$$

Rearranging yields the Bregman bound

$$A_x(\theta') - A_x(\theta) - \langle \nabla A_x(\theta),\, \theta' - \theta\rangle \le \frac{1}{2}\|\theta' - \theta\|_2^2.$$

Thus

$$D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x) \,\|\, \pi_{\theta'}(\cdot|x)\big) \le \frac{1}{2}\|\theta - \theta'\|_2^2.$$

Substituting $(\theta, \theta') = (\theta_T^{\mathrm{atk}}, \theta_T^{\mathrm{cln}})$ proves the first claim, and taking expectation over x preserves the inequality. Finally, apply Theorem 3.2 to bound $\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2$. □

F.11 Proof of Theorem 3.4

Proof. For any attack set A with $|A| \le K$,

$$\nabla \tilde{F}_{\tau,A}(\theta) - \nabla F(\theta) = \big(\nabla F_\tau(\theta) - \nabla F(\theta)\big) + \frac{1}{W_\tau}\sum_{i\in A} w_i^{(\tau)} \beta v_i.$$

The first term is bounded by $\mathsf{B}_\tau^{\mathrm{cln}}(R)$ by definition. For the second term, using $\|\beta v_i\|_2 = s_i$ gives

$$\left\| \frac{1}{W_\tau}\sum_{i\in A} w_i^{(\tau)} \beta v_i \right\|_2 \le \frac{1}{W_\tau}\sum_{i\in A} w_i^{(\tau)} s_i = \frac{1}{W_\tau}\sum_{i\in A} \rho_i^{(\tau)} \le \frac{\Gamma_\tau(K)}{W_\tau},$$

where the last inequality follows from the definition of $\Gamma_\tau(K)$ as the largest possible sum of K weighted impacts. This proves (14).

For the exponential weights, every original top-K sample has weight at most $\varepsilon_\tau = e^{-1/\tau}$, and every sample outside the original top-K has score at most $\delta(K)$. Therefore, for any A with $|A| \le K$,

$$\sum_{i\in A} \rho_i^{(\tau)} \le \varepsilon_\tau \sum_{j=1}^{K} s_{(j)} + K\delta(K).$$

Taking the maximum over A proves the upper-bound for $\Gamma_\tau(K)$. □

F.12 Proof of Theorem 3.5

Proof. The Hessian of $\tilde{F}_{\tau,A}$ is the same as the Hessian of F_τ, because label swaps leave each matrix $\sigma(z_i)(1-\sigma(z_i))v_iv_i^\top$ unchanged. The weighted clean objective satisfies

$$\nabla^2 F_\tau(\theta) = \frac{1}{W_\tau}\sum_{i=1}^{N} w_i^{(\tau)}\beta^2\sigma(z_i(\theta))\big(1-\sigma(z_i(\theta))\big)v_iv_i^\top.$$

By the definitions of $\underline{c}_\tau(R)$ and Σ_τ,

$$\nabla^2 F_\tau(\theta) \succeq \beta^2\underline{c}_\tau(R)\Sigma_\tau \succeq \mu_\tau I$$

for all $\theta \in \mathbb{B}_R(\theta_{\mathrm{ref}})$. Thus $\tilde{F}_{\tau,A}$ is μ_τ-strongly convex on the bounded region and is β^2-smooth by Lemma E.1. Theorem 3.4 gives

$$\sup_{\theta\in\mathbb{B}_R(\theta_{\mathrm{ref}})} \|\nabla\tilde{F}_{\tau,A}(\theta) - \nabla F(\theta)\|_2 \le \mathsf{B}_\tau^{\mathrm{cln}}(R) + \frac{\Gamma_\tau(K)}{W_\tau}.$$

Applying Lemma E.6 with $F_0 = F$, $H = \tilde{F}_{\tau,A}$, $\mu = \mu_\tau$, $L = \beta^2$, and

$$B = \mathsf{B}_\tau^{\mathrm{cln}}(R) + \frac{\Gamma_\tau(K)}{W_\tau}$$

proves the first displayed bound. The second follows by substituting the coarse bound (15). □

F.13 Proof of Corollary 3.6

Proof. The KL-to-parameter conversion is the same as in the proof of Corollary 3.3. Applying Theorem 3.5 gives the stated bound. □

F.14 Proof of Proposition 3.7

Proof. On the bounded region, $F_{\mathfrak{R}}$ is $\mu_{\mathfrak{R}}$-strongly convex and β^2-smooth by assumption. Adding the quadratic regularizer adds Hessian λI. Hence

$$\nabla^2 F_{\mathfrak{R},\lambda}(\theta) = \nabla^2 F_{\mathfrak{R}}(\theta) + \lambda I \succeq (\mu_{\mathfrak{R}} + \lambda)I,$$

and

$$\|\nabla^2 F_{\mathfrak{R},\lambda}(\theta)\|_{\mathrm{op}} \le \|\nabla^2 F_{\mathfrak{R}}(\theta)\|_{\mathrm{op}} + \lambda \le \beta^2 + \lambda.$$

This proves $(\mu_{\mathfrak{R}} + \lambda)$-strong convexity and $(\beta^2 + \lambda)$-smoothness. □

F.15 Proof of Corollary 3.8

Proof. Fix $\mathfrak{R} \in \{G, \tau\}$ and an admissible attack set A. Assumption 2.3 supplies the bounded training region and curvature on the directions updated by both runs. By Lemma 2.2, label flips change the gradient of either fixed robust objective by a constant vector. Consequently, $\tilde{F}_{\mathfrak{R},\lambda,A}$ has the same Hessian as $F_{\mathfrak{R},\lambda}$ and, by Proposition 3.7, is $(\mu_{\mathfrak{R}} + \lambda)$-strongly convex and $(\beta^2 + \lambda)$-smooth on that region, in the trained coordinates of Assumption 2.3. The regularizer cancels from the attacked-to-clean robust difference, giving

$$\nabla\tilde{F}_{\mathfrak{R},\lambda,A}(\theta) - \nabla F(\theta) = \big(\nabla F_{\mathfrak{R},\lambda}(\theta) - \nabla F(\theta)\big) + \big(\nabla\tilde{F}_{\mathfrak{R},A}(\theta) - \nabla F_{\mathfrak{R}}(\theta)\big).$$

The first term has norm at most $\mathsf{B}^{\mathrm{cln}}_{\mathfrak{R},\lambda}(R)$. The second has norm at most $\Delta_{\mathfrak{R}} - \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R)$: this is $K\delta(K)/(N-K)$ by the proof of Theorem 3.1 when $\mathfrak{R} = G$, and $\Gamma_\tau(K)/W_\tau$ by the proof of Theorem 3.4 when $\mathfrak{R} = \tau$. The common initialization and Assumption 2.3 place both trajectories in the same affine training space. Restricting to its coordinates preserves the updates and cannot increase the gradient-perturbation bound. Apply Lemma E.6 in those coordinates with $F_0 = F$, $H = \tilde{F}_{\mathfrak{R},\lambda,A}$, $\mu = \mu_{\mathfrak{R}} + \lambda$, $L = \beta^2 + \lambda$, and

$$B = \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R},\lambda}(R) + \Delta_{\mathfrak{R}} - \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R)$$

to obtain the first inequality. For the second, use $\mathsf{B}^{\mathrm{cln}}_{\mathfrak{R},\lambda}(R) \le \mathsf{B}^{\mathrm{cln}}_{\mathfrak{R}}(R) + \lambda R$ and $\Delta_{\mathfrak{R}} \le \bar{\Delta}_{\mathfrak{R}}$. The latter follows from Lemma E.3 for G, and Theorem 3.4 together with Lemma E.4 for τ. □

F.16 Proof of Corollary 3.9

Proof. Fix either $\mathfrak{R} \in \{G, \tau\}$. As in the proof of Corollary 3.6, Assumption 2.1 implies that the log-partition function is 1-smooth. Its Bregman-divergence representation of conditional KL therefore gives, for every prompt x,

$$D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big) \le \frac{1}{2}\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2^2.$$

The parameter-distance term is independent of x. Taking expectation over prompts and applying Corollary 3.8 yields

$$\mathbb{E}_x\Big[D_{\mathrm{KL}}\big(\pi_{\theta_T^{\mathrm{atk}}}(\cdot|x) \,\|\, \pi_{\theta_T^{\mathrm{cln}}}(\cdot|x)\big)\Big] \le \frac{1}{2}\|\theta_T^{\mathrm{atk}} - \theta_T^{\mathrm{cln}}\|_2^2 \le \frac{1}{2}\left(\frac{\bar{\Delta}_{\mathfrak{R}} + \lambda R}{\mu_{\mathfrak{R}} + \lambda}\right)^2,$$

which proves (17). □

G Experiments: Additional details for the SHP simulation

Data and feature construction. The SHP simulation uses 200,000 decoded preference pairs from the SHP training split (seed 0), followed by an $80/20$ train/test split that yields 160,000 training pairs and 40,000 held-out test pairs. For each example, we use the feature difference v_i defined in (3). We concatenate the prompt and response with two newline characters, tokenize with a maximum length of 256 tokens, and extract the final-layer CLS representation from `sentence-transformers/all-MiniLM-L6-v2`. We ℓ_2-normalize this vector to obtain $\psi(x, y) \in \mathbb{R}^{384}$, so $\|v_i\|_2 \le 2$. This instantiates the log-linear DPO parameterization. SHP is a public Reddit-derived preference dataset. We cite its associated paper and follow the dataset-card terms. The embedding model `sentence-transformers/all-MiniLM-L6-v2` is released under the Apache-2.0 license.

Reference policy. The reference parameter $\theta_{\mathrm{ref}} \in \mathbb{R}^{384}$ is trained *independently of the DPO runs* on the same training split using a pairwise logistic ranking loss

$$\mathcal{L}_{\mathrm{ref}}(\theta) \;=\; \mathbb{E}_i\big[\log\big(1 + \exp(-\langle\theta, v_i\rangle)\big)\big] \;+\; \tfrac{\lambda_{\mathrm{ref}}}{2}\,\|\theta\|_2^2,$$

with $\lambda_{\mathrm{ref}} = 10^{-3}$, 5,000 mini-batch SGD steps, learning rate 0.25, and batch size 100,000. The resulting parameter has $\|\theta_{\mathrm{ref}}\|_2 \approx 2.26$. The reference parameter defines the reference-policy log-ratio in the DPO loss, provides the common initialization, and anchors the quadratic regularizer. It is not treated as a near-optimal policy. All DPO and robust-DPO runs are initialized at $\theta_0 = \theta_{\mathrm{ref}}$.

Optimization and attack configuration. All models are trained by SGD with learning rate 0.25, batch size 25,600, and 100,000 optimization steps, with full-batch evaluation every 100 steps. We set $\beta = 1$ and attack fraction 0.3, giving a nominal attacked-scenario budget of $K = \lfloor 0.3 \cdot 160{,}000 \rfloor = 48{,}000$ flipped training pairs. Four attack settings are considered:

- **No attack**: the training data is clean.
- **Random corruption**: K training pairs are randomly selected and flipped.
- **Top-impact attack**: K training pairs with the largest impact scores $s_i = \beta\|v_i\|_2$ are flipped.
- **Targeted BMP-A attack** (Yang et al., 2026): K training pairs are selected by BMP-A, a gradient-space label-flip attack that drives DPO toward the target policy $\theta^\dagger = 2\theta_{\mathrm{ref}} - \theta^\star$. The detailed procedure and construction of this adversarial attack are provided in Appendix H.

For the attacked scenarios, `TRIM` and `TRIM_REG` remove the top-K examples under the observed impact ranking ($K = 48{,}000$, leaving 112,000 training pairs). The REG variants use $\lambda = 10^{-3}$. For the no-attack control, we set $K = 0$ while keeping $\lambda = 10^{-3}$ for the REG controls. The unregularized proposed defenses then reduce to standard DPO, and the REG variants reduce to full-data regularized DPO. For `SOFT_TRIM` and `SOFT_TRIM_REG`, the full training set is retained with exponential weights at $\tau = 20{,}000$. When $K = 0$, all weights are one. We also run a clean-data wrong-K stress test in which the realized attack budget is zero but the defender uses the nominal attacked-scenario budget $K = 48{,}000$. This measures the clean-data cost of overestimating K.

Compute resources. The reported runs used Apple Silicon M4 Max with MPS for sequential preprocessing, clean-reference training, and targeted-attack construction. The remaining algorithm cases were dispatched to CPU worker processes. The logged run used 64GB unified memory, 12 performance cores, 4 efficiency cores, 14 parallel workers, and finished the full no-attack, random, top-impact, and targeted-BMP-A sweep in about two hours.

Compared models and legend names. We use the shorthand defined in Section 4: `DPO`, `DPO_REG`, `TRIM`, `TRIM_REG`, `SOFT_TRIM`, and `SOFT_TRIM_REG`. The additional baselines are `DR_DPO`, `NR_DPO`, `WDPO`, and `KLDPO` (Appendix J). In the conservative wrong-K scenario, "DPO (trimmed data)" denotes DPO trained on the retained clean subset. In the ordinary no-attack scenario ($K = 0$), no pairs are removed and the legend uses `DPO`. The matched full-data `DPO_REG` control uses the same observed labels and $\lambda = 10^{-3}$ without trimming or downweighting (Appendix K). The clean full-set DPO reference is displayed as $\pi^\star$ in policy KL and as $\theta^\star$ in parameter L_2 plots.

Evaluation metrics. Our primary metric is the *candidate-set policy distance*, defined as the average categorical KL between two induced finite-action policies on held-out prompts. For each evaluation prompt x, we construct an indexed candidate list of size $M = 100$:

$$\mathcal{Y}(x) = (y_w, y_l, y_{\text{neg}}^{(1)}, \ldots, y_{\text{neg}}^{(M-2)}).$$

The response pool contains both responses from each of the 160,000 training pairs, with duplicate and empty strings removed. We sample $M - 2 = 98$ additional responses with replacement using seed 123, excluding strings equal to y_w or y_l. Repeated sampled responses remain separate candidate entries. Writing $\mathcal{Y}_j(x)$ for entry j, the evaluated categorical policy is

$$\pi_\theta(j \mid x; \mathcal{Y}(x)) = \frac{\exp(\langle \theta, \psi(x, \mathcal{Y}_j(x)) \rangle)}{\sum_{k=1}^{M} \exp(\langle \theta, \psi(x, \mathcal{Y}_k(x)) \rangle)}.$$

The same candidate lists are fixed across all compared methods and checkpoints. We then report

$$\text{candKL}(\theta) \;:=\; \mathbb{E}_x[\mathrm{KL}(\pi^\star(\cdot \mid x; \mathcal{Y}(x)) \,\|\, \pi_\theta(\cdot \mid x; \mathcal{Y}(x)))]$$

averaged over 5,000 held-out prompts. We also record parameter distance $\|\theta - \theta^\star\|_2$, but candKL is the primary metric as it measures the downstream policy deviation on a richer action space, and it reflects the effect of attack and defense more directly.

Candidate-set KL measures deviation from the clean-DPO policy on the fixed evaluation candidates. It does not directly measure absolute response quality.

H Experiments: BMP-A Targeted Attack: Full Construction

This appendix explains how we instantiate the Binary Matching Pursuit Attack (BMP-A) from Yang et al. (2026) in our robust DPO experiments. BMP-A is used only to generate the attacked training set. All robust defense methods then train on this same attacked dataset or on the corresponding trimmed subset. BMP-A selects flips to match an attacker-chosen target gradient, whereas top-impact selects by magnitude and random corruption selects uniformly. Their realized policy effects depend on the target, budget, and data geometry.

Gradient-space reduction. Recall the feature difference v_i and affine logit $z_i(\theta)$ from (3). In our SHP simulation, the reference policy is parameterized by θ_{ref}, so $b_i = \langle \theta_{\text{ref}}, v_i \rangle$ and the logit specializes to

$$z_i(\theta) = \beta \langle \theta - \theta_{\text{ref}}, v_i \rangle.$$

By Lemma 2.2, flipping pair i changes the clean per-sample gradient $g_i(\theta)$ by the fixed vector βv_i. Thus a support A changes the unnormalized gradient sum by $\sum_{i \in A} \beta v_i$, or the gradient of the averaged objective F by $N^{-1} \sum_{i \in A} \beta v_i$. This is the key observation used in Yang et al. (2026): targeted label flipping can be formulated as a sparse binary approximation problem in gradient space.

Target policy. We choose the adversarial target as the reflection of the clean DPO solution across the reference parameter,

$$\theta^\dagger = 2\theta_{\mathrm{ref}} - \theta^\star.$$

This target asks the attacked optimizer to move away from θ_{ref} by roughly the same distance as the clean optimizer, but in the opposite direction. Define the unnormalized clean-gradient sum and its cancelling target vector by

$$g^\dagger := \sum_{i=1}^{N} g_i(\theta^\dagger), \qquad y := -g^\dagger.$$

If the attacker can find a flip set A such that $\sum_{i\in A} \beta v_i \approx y$, then the attacked DPO gradient at $\theta^\dagger$ is close to zero. Thus $\theta^\dagger$ becomes an approximate stationary point of the attacked objective.

Sparse binary attack problem. Let

$$V_{\mathrm{atk}} := [\beta v_1, \ldots, \beta v_N] \in \mathbb{R}^{d\times N}.$$

The attacker seeks a binary selector $x \in \{0,1\}^N$ with $\|x\|_0 \le K$ such that

$$V_{\mathrm{atk}} x \approx y. \tag{21}$$

Here $x_i = 1$ means that pair i is flipped. This is the same gradient-space attack formulation as in Yang et al. (2026), specialized to the target $\theta^\dagger$ used in our experiments.

BMP-A selection rule. BMP-A is a greedy solver for (21). Starting from residual $r = y$ and empty support A, it repeatedly selects an unflipped pair whose normalized column direction has the largest correlation with the current residual:

$$i_t \in \arg\max_{i\in[N]\setminus A} \frac{|\langle V_{\mathrm{atk},i}, r\rangle|}{\max\{\|V_{\mathrm{atk},i}\|_2, 10^{-12}\}}.$$

The normalization is important because the columns βv_i are not unit-norm. The numerical floor matches the implementation and gives zero score to zero-impact columns. After choosing i_t, BMP-A sets $x_{i_t} = 1$ and updates

$$r \leftarrow r - V_{\mathrm{atk},i_t},$$

so that the invariant $r = y - V_{\mathrm{atk}} x$ is maintained. The procedure stops when the budget K is exhausted or when the residual norm is below the tolerance. Algorithm 3 gives the exact procedure used in our experiments.

Algorithm 3 Binary Matching Pursuit Attack (BMP-A) used for targeted label flipping

Require: Attack matrix $V_{\mathrm{atk}} = [\beta v_1, \ldots, \beta v_N]$, target $y = -g^\dagger$, budget K, tolerance ε
Ensure: Binary flip vector $x \in \{0,1\}^N$, flip set A

$r \leftarrow y, \quad x \leftarrow 0, \quad A \leftarrow \emptyset$
for $t = 1, 2, \ldots, K$ **do**
  $i_t \in \arg\max_{i\in[N]\setminus A} \dfrac{|\langle V_{\mathrm{atk},i}, r\rangle|}{\|V_{\mathrm{atk},i}\|_2}$
  $A \leftarrow A \cup \{i_t\}, \quad x_{i_t} \leftarrow 1$
  $r \leftarrow r - V_{\mathrm{atk},i_t}$
  if $\|r\|_2 \le \varepsilon$ **then**
    break
  end if
end for

Forming the attacked dataset. After BMP-A returns the support A, we form $\tilde{\mathcal{D}}(A)$ using the label-flip rule in Section 2.2. This attacked dataset is fixed before defense training and shared by all methods.

Target-strength line search. A fixed budget may not be sufficient to reach the full opposite target. Therefore, our implementation also considers intermediate targets

$$\theta^\dagger(\alpha) = \theta_{\text{ref}} + \alpha(\theta^\dagger - \theta_{\text{ref}}), \qquad \alpha \in (0, 1].$$

For each α, we compute $g^\dagger(\alpha) = \sum_i g_i(\theta^\dagger(\alpha))$, set $y(\alpha) = -g^\dagger(\alpha)$, and run BMP-A with budget K. We scan α from large to small over $\{1.0, 0.9, \ldots, 0.1\}$ and use the first value satisfying

$$\frac{\|y(\alpha) - V_{\text{atk}} x\|_2}{\|y(\alpha)\|_2} \le 0.15.$$

If no value meets this criterion, we use the α with the smallest relative residual. This selects the largest tested α meeting the residual criterion when one is available.

Realized BMP-A instance. In our SHP experiment, $N = 160{,}000$, $d = 384$, $K = 48{,}000$, $\beta = 1$, and the residual tolerance is $\varepsilon = 10^{-3}$. The line search accepts $\alpha = 1.0$, so the full target $\theta^\dagger = 2\theta_{\text{ref}} - \theta^\star$ is used. The final BMP-A support has size $48{,}000$, the relative gradient-space residual is 0.094, and

$$\|\theta^\dagger - \theta^\star\|_2 \approx 40.3.$$

Training standard DPO on this attacked dataset yields a final parameter distance around 30.7 from $\theta^\star$. The constructed attack therefore causes a substantial deviation from clean DPO, providing a targeted-attack setting for evaluating the defenses.

I Experiments: No-Attack and Random-Corruption Results

This appendix reports results for the no-attack, clean-data wrong-K, and random-corruption scenarios, and provides parameter-distance ($\|\theta - \theta^\star\|_2$) curves for all four attack scenarios. The figures in this section compare standard DPO with the four proposed defenses (`TRIM`, `TRIM_REG`, `SOFT_TRIM`, and `SOFT_TRIM_REG`). Comparisons with the four additional baselines appear in Appendix J.

I.1 No-attack scenario

When the training data are clean, the realized attack budget is $K = 0$. With no samples to trim and all soft-trim weights equal to one, `TRIM` and `SOFT_TRIM` reduce exactly to standard DPO. The `_REG` controls keep $\lambda = 10^{-3}$, so `TRIM_REG` and `SOFT_TRIM_REG` reduce to full-data regularized DPO and retain a clean-data offset from the unregularized reference $\pi^\star$. Figure 3 confirms that both unregularized defenses retain the clean-DPO solution when $K = 0$, with final candKL $= 0$ up to numerical tolerance. The REG controls isolate the deviation introduced by the quadratic regularizer.

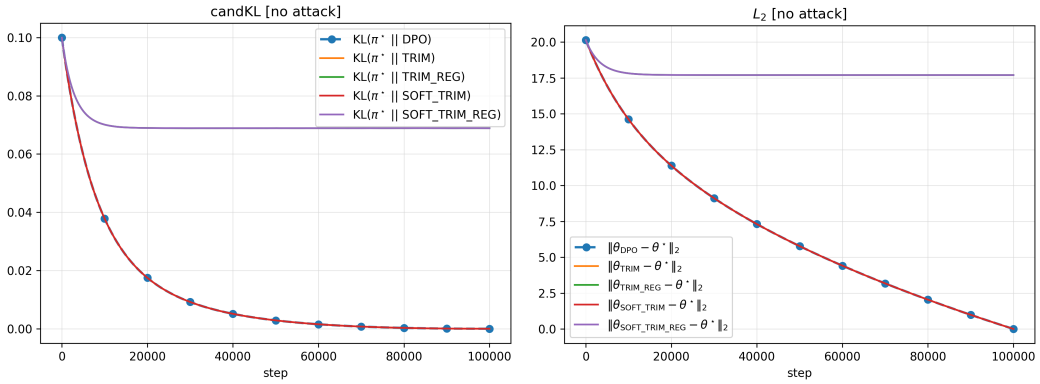


Figure 3: No-attack scenario. candKL, plotted as $\mathrm{KL}(\pi^\star\|\cdot)$ (left), and $\|\theta - \theta^\star\|_2$ (right). The unregularized proposed defenses use $K = 0$, while the REG variants keep $\lambda = 10^{-3}$.

I.2 NO-ATTACK SCENARIO WITH CONSERVATIVE WRONG K

We also train on clean data while setting the trimming budget to $K = 48{,}000$. Hard trimming removes 30% of the clean training set, giving final candKL 0.018 for `TRIM`. Soft trimming retains all pairs and reaches candKL 0.007.

Figure 4 illustrates the clean-data cost of overestimating K. `SOFT_TRIM` incurs a smaller deviation than `TRIM` in this setting, consistent with its less aggressive downweighting rule.

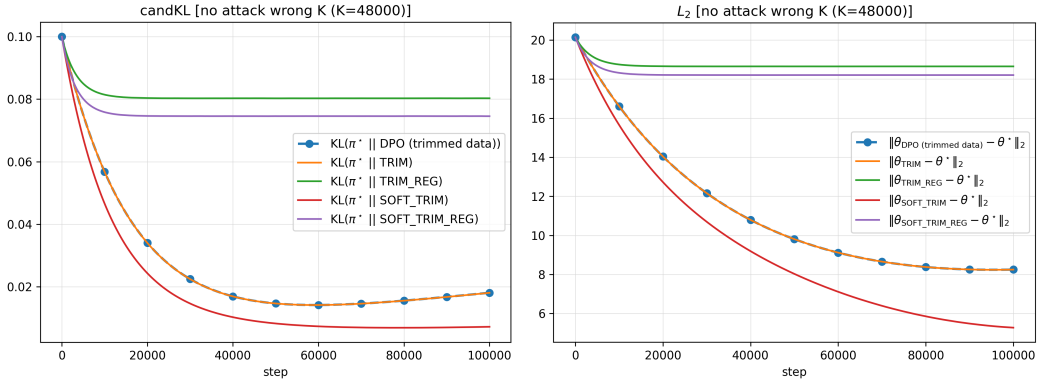


Figure 4: No-attack scenario with conservative wrong K. candKL, plotted as $\mathrm{KL}(\pi^\star\|\cdot)$ (left), and $\|\theta - \theta^\star\|_2$ (right). The true attack budget is zero, but the defender uses $K = 48{,}000$, so hard trimming discards clean high-impact pairs and soft trimming keeps a smaller clean-data bias.

I.3 RANDOM-CORRUPTION SCENARIO

Under 30% random label flips, only $\approx 30\%$ of the top-K impact samples are actually corrupted, so `TRIM`'s precision and recall for identifying attacked samples are both ≈ 0.30. Across ten random seeds, standard `DPO` reaches candKL $= 0.041 \pm 0.001$, compared with 0.039 ± 0.003 for `TRIM` and 0.037 ± 0.002 for `SOFT_TRIM`. The regularized variants have higher final candKL, with 0.0918 ± 0.0004 for `TRIM_REG` and 0.0892 ± 0.0003 for `SOFT_TRIM_REG`. The $\pm$ values are empirical standard deviations across random attack seeds. Figure 5 shows modest reductions in mean candKL for the unregularized defenses, even though the impact ranking does not preferentially select randomly flipped pairs. Regularization does not improve this comparison, consistent with the clean-data offset observed in the no-attack control.

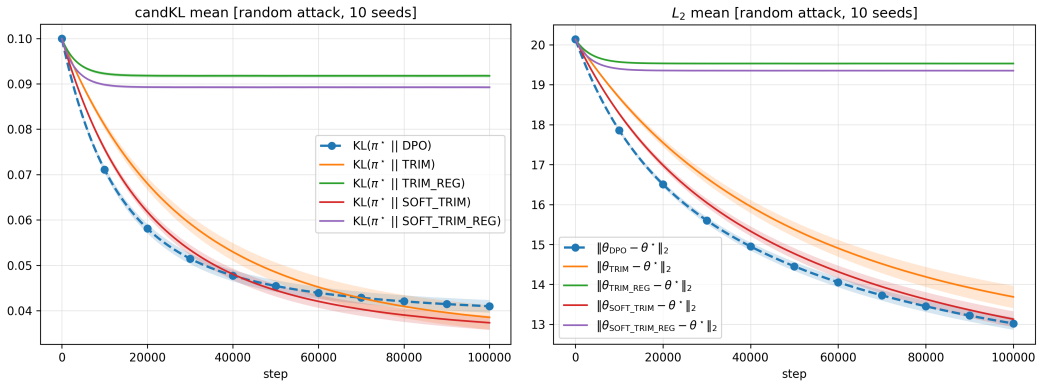


Figure 5: Random-corruption scenario (30% of pairs flipped uniformly at random), averaged over ten seeds. Left: candKL from clean DPO. Right: parameter distance from $\theta^\star$.

I.4 Parameter-distance curves for top-impact and targeted attacks

For completeness, Figures 6 and 7 show the $\|\theta - \theta^\star\|_2$ trajectories for the top-impact and targeted attacks. Under the top-impact attack, TRIM drives θ-distance to 8.25 (identical to TRIM on clean data with the same K), while attacked DPO reaches 25.2. Under the targeted BMP-A attack, TRIM_REG ($\|\theta - \theta^\star\|_2 = 21.2$) and SOFT_TRIM_REG (21.5) are substantially less deflected than unregularized TRIM (32.3) and DPO (30.7), consistent with the candKL results.

The parameter-distance comparisons support the same conclusions as candKL: hard trimming is effective against the top-impact attack, while the REG variants reduce deviation from clean DPO under BMP-A.

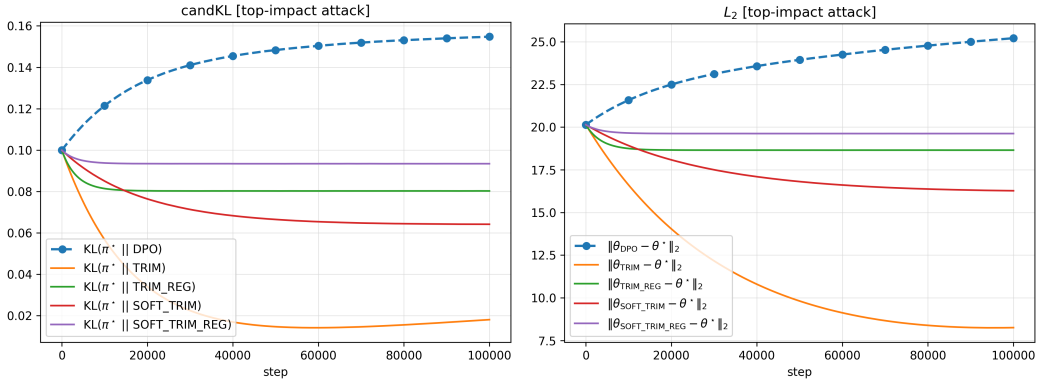


Figure 6: Top-impact attack. candKL, plotted as $\mathrm{KL}(\pi^\star\|\cdot)$ (left), and $\|\theta - \theta^\star\|_2$ (right). TRIM matches its clean-data counterpart with the same K.

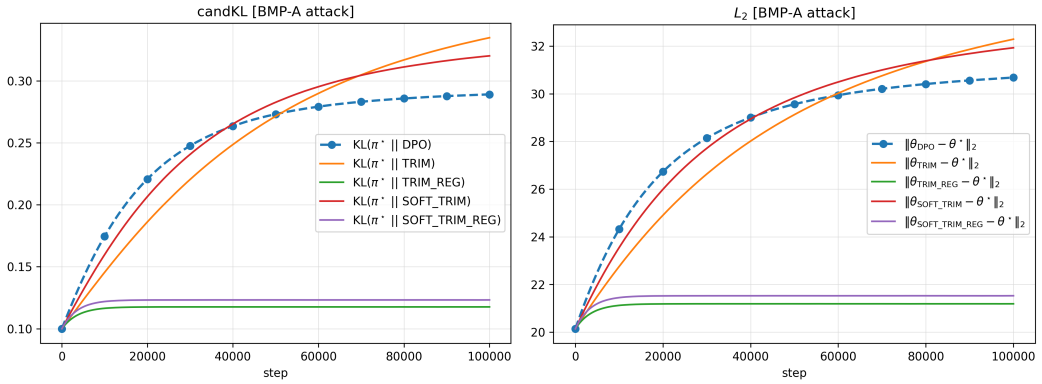


Figure 7: Targeted BMP-A attack. candKL, plotted as $\mathrm{KL}(\pi^\star\|\cdot)$ (left), and $\|\theta - \theta^\star\|_2$ (right). Both REG variants reduce deviation from clean DPO.

J Experiments: Comparison with DR_DPO, NR_DPO, WDPO, and KLDPO Baselines

We compare our proposed defenses against four existing robust DPO baselines designed for random noise or distributional shift.

J.1 Baseline methods

DR_DPO (Wu et al., 2024b). DR_DPO applies an entropic aggregation to the DPO log-likelihood over preference pairs. In our positive-loss notation, the implemented objective is

$$\mathcal{L}_{\mathrm{DR}}(\pi_\theta) \;=\; -\beta' \log \mathbb{E}\Big[\exp\Big(-\tilde{\ell}_i(\theta)/\beta'\Big)\Big]\,,$$

where $\beta' > 0$ is a tunable temperature. We use $\beta' = 1.0$. The original DR_DPO objective applies the entropic transform to $h_i = -\tilde{\ell}_i$, the DPO log-likelihood, following Appendix D of Wu et al. (2024b). In positive-loss notation, this gives a soft-minimum rather than the KL-DRO worst-case risk $\beta' \log \mathbb{E}[\exp(\tilde{\ell}_i/\beta')]$. The induced per-sample gradient weight is $\exp(-\tilde{\ell}_i/\beta')/Z$, which down-weights pairs with high observed loss. This weighting can reduce the influence of high-loss noisy pairs, but high loss can also reflect a clean, difficult comparison. Under targeted corruption, fitting attacked labels can make opposing clean comparisons harder, so this weighting need not suppress the poisoned signal.

NR_DPO (Chowdhury et al., 2024). NR_DPO (called rDPO in the original paper) applies the symmetric-flip unbiased correction:

$$\ell_{\mathrm{NR}}(\theta; z) \;=\; \frac{(1-\epsilon)\,\ell(z) - \epsilon\,\ell(-z)}{1-2\epsilon},$$

where z is the chosen-rejected logit and $\epsilon \in [0, 1/2)$ is the assumed flip rate. We set ϵ equal to the true attack fraction $= 0.3$ in attack scenarios, and to 0 in the no-attack scenario, which recovers standard DPO. This correction is unbiased for the clean DPO loss under i.i.d. symmetric label flips, and would work well when corruption matches that model.

WDPO (Xu et al., 2025). WDPO targets distributional shift in preference data, e.g. across demographic or temporal subpopulations, via a Wasserstein-DRO formulation $\min_\theta \sup_{P:\, W_p(P,P_o)\le\rho} \mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)]$. Following Algorithm 1 of Xu et al. (2025), we tractably approximate this by adding an input-gradient penalty to the empirical DPO loss. In the SHP simulator, the observed pair has $\tilde{v}_i = \psi(x^i, \tilde{y}_w^i) - \psi(x^i, \tilde{y}_l^i)$ and logit $\tilde{z}_i(\theta) = \beta\langle\theta - \theta_{\mathrm{ref}}, \tilde{v}_i\rangle$. The implementation adds $2\rho \cdot \mathbb{E}[\beta^2\sigma(-\tilde{z}_i(\theta))^2\|\theta - \theta_{\mathrm{ref}}\|_2^2]$ to the empirical loss. The factor 2 accounts for the sum of squared gradient norms with respect to the chosen and rejected feature vectors, since their difference is $\tilde{v}_i$. This is the log-linear analogue of the embedding-gradient penalty in WDPO. We set $\rho = 10^{-2}$ for our simulator.

KLDPO (Xu et al., 2025). KLDPO targets the same distributional-shift setting as WDPO but with a KL-divergence uncertainty set, leading to a worst-case kernel that upweights samples with above-average loss. Our minibatch implementation uses $w_i = \min\{10, \exp((\tilde{\ell}_i - \bar{\tilde{\ell}})/\tau_{\mathrm{kl}})\}$ with temperature $\tau_{\mathrm{kl}} = 0.05$, where $\bar{\tilde{\ell}}$ is the current minibatch mean loss. These clipped weights are held fixed when differentiating the minibatch mean of $w_i\tilde{\ell}_i$. This is the approximate reweighting update used in our comparison with Algorithm 2 of Xu et al. (2025).

J.2 Results and discussion

We evaluate all four robust-DPO baselines together with our four proposed defenses (TRIM, TRIM_REG, SOFT_TRIM, SOFT_TRIM_REG) on the no-attack, random-corruption, top-impact, and targeted-BMP-A scenarios introduced in Section 4. Per-step candKL and parameter-distance trajectories for each scenario are shown in Figures 8–11. The final-step candKL values are summarized in Table 1. For the random-corruption column, all methods are averaged over the same ten random attack seeds. The proposed-method means and standard deviations are reported separately in Appendix I.

No-attack scenario (Figure 8). When the training data are clean, the proposed TRIM and SOFT_TRIM defenses are evaluated with $K = 0$, so the unregularized variants coincide with standard DPO and have final candKL $= 0$. The REG variants keep $\lambda = 10^{-3}$ and therefore report the clean-data offset of full-data regularized DPO. NR_DPO (with $\epsilon = 0$, recovering standard DPO) and DR_DPO also converge to near-zero candKL. WDPO (candKL $= 0.096$) and KLDPO (candKL $= 0.089$) remain farther from the full-data clean-DPO reference. Their penalty and reweighting rules change the training updates even when no corruption is present.

Thus, the unregularized TRIM and SOFT_TRIM methods preserve clean-DPO training when the defender sets $K = 0$. The REG controls retain the bias introduced by their quadratic penalty.

Random corruption (Figure 9). Under 30% uniformly random label flips, the corruption setting is closest to the noise model used by NR_DPO. NR_DPO (noise rate $\epsilon = 0.3$) reaches the lowest average

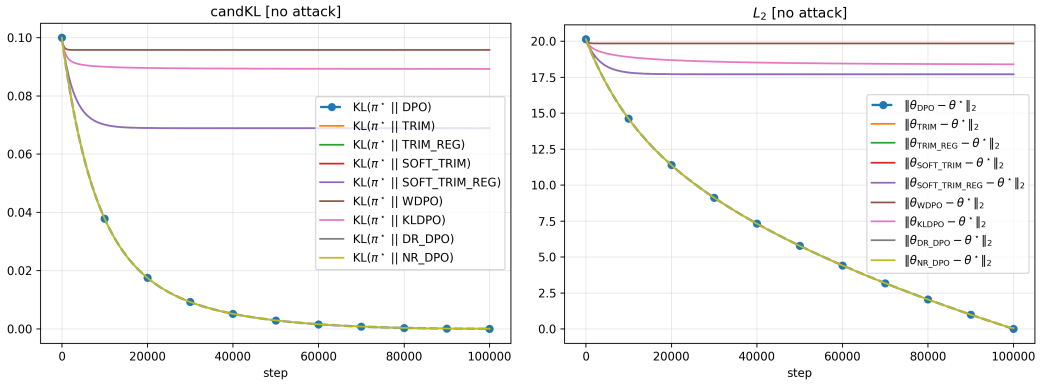


Figure 8: No-attack scenario with all baselines. Left: candKL from clean DPO. Right: parameter distance from $\theta^\star$. The proposed defenses use $K = 0$, and the REG variants keep $\lambda = 10^{-3}$. NR_DPO (with $\epsilon = 0$) and DR_DPO also converge to near-zero candKL. WDPO and KLDPO converge to a moderately different solution (candKL ≈ 0.09–0.10).

candKL among the compared methods under random flips: candKL $= 0.010$, approximately one quarter of standard DPO's 0.041. This is consistent with its random-noise correction. DR_DPO (0.082), WDPO (0.098), and KLDPO (0.096) have higher candKL than standard DPO.

The unregularized proposed methods give modest improvements over DPO: the ten-seed mean candKL values are 0.039 for TRIM and 0.037 for SOFT_TRIM, compared with 0.041 for DPO.

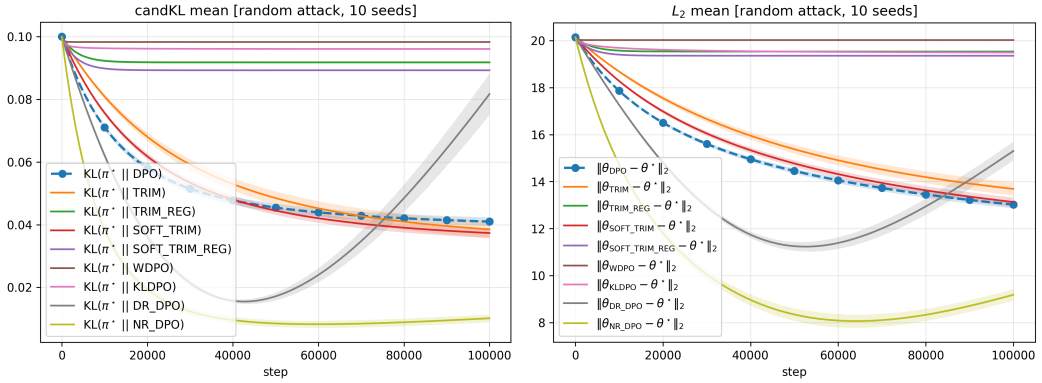


Figure 9: Random-corruption scenario with all baselines, averaged over ten seeds. Left: candKL from clean DPO. Right: parameter distance from $\theta^\star$.

Top-impact attack (Figure 10). Under the top-impact attack, DR_DPO (candKL $= 0.423$) and NR_DPO (0.304) deviate more than attacked DPO (0.155). WDPO and KLDPO have lower candKL than DPO (0.101 and 0.102), but remain above TRIM (0.018). SOFT_TRIM also has lower candKL (0.064) than these baselines.

Both unregularized proposed methods outperform the compared baselines in this setting. Hard trimming removes all flipped pairs and matches TRIM trained on clean data with the same K, leaving only the deviation due to discarding clean training information.

Targeted BMP-A attack (Figure 11). Under the targeted BMP-A attack, NR_DPO (candKL $=$ 1.079) and DR_DPO (1.066) have larger deviations than standard DPO (0.289). Their parameter distances are 54–58, compared with 30.7 for DPO. WDPO (candKL $= 0.103$) and KLDPO (candKL $= 0.107$) have smaller deviations than DPO in this fixed BMP-A instance. TRIM (candKL $= 0.335$) and SOFT_TRIM (candKL $= 0.320$) are worse than standard DPO in this instance. Their top-K intervention region overlaps only 42% of the corrupted support: hard trim-

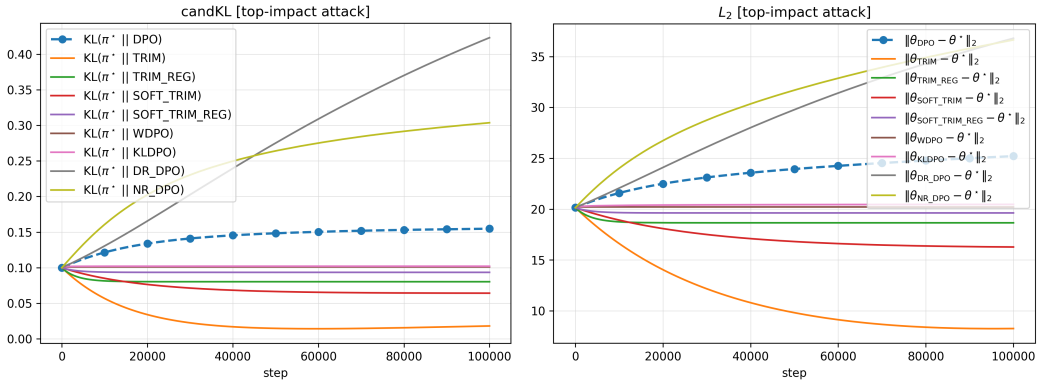


Figure 10: Top-impact attack with all baselines. Left: candKL from clean DPO. Right: parameter distance from $\theta^\star$.

ming removes those pairs, while soft trimming downweights them. Both rules also suppress clean high-impact samples. The regularized variants are closer to full-data clean DPO, with candKL 0.118 for TRIM_REG and 0.123 for SOFT_TRIM_REG, both below matched full-data DPO_REG (0.130).

Here regularization helps the proposed defenses reduce policy deviation under BMP-A, while unregularized trimming alone is insufficient. Both proposed REG variants improve over DPO, but WDPO and KLDPO achieve lower candKL in this instance. Appendix K quantifies the additional benefit of trimming beyond the shared regularizer.

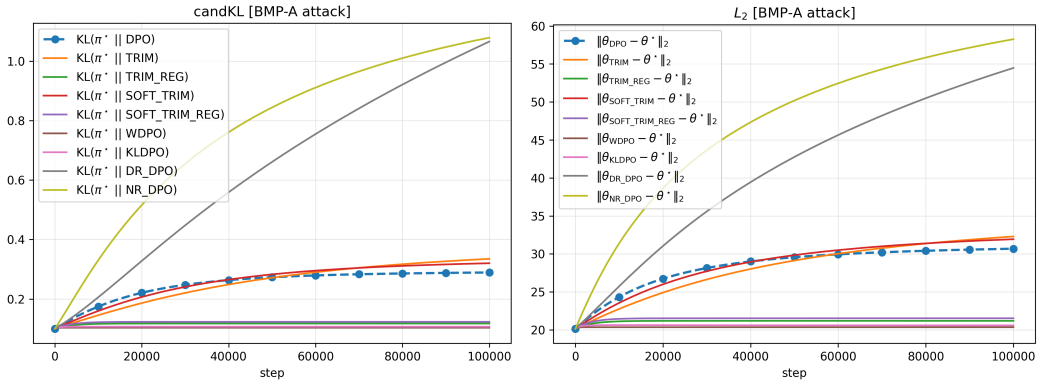


Figure 11: Targeted BMP-A attack with all baselines. Left: candKL from clean DPO. Right: parameter distance from $\theta^\star$.

Table 1 summarizes the final-step candKL across all four scenarios.

For NR_DPO, the strong performance under random flips does not carry over to targeted BMP-A. The independent random-noise model required for NR_DPO's cancellation is violated by this fixed targeted support. DR_DPO weights easy observed pairs more strongly. If those are attacked, the weighting can favor the poisoned signal. These mechanisms explain why their respective robustness assumptions need not hold here.

WDPO and KLDPO stay in a narrow range (candKL $\in [0.09, 0.11]$ in these runs). This makes them relatively stable under the targeted BMP-A attack, where WDPO has the lowest final candKL in this run. Their no-attack deviations also show that the modified training rules do not recover the full-data clean-DPO reference in these runs. Their design addresses distribution shift rather than identifying and removing a label-flipped subset.

These comparisons support the effectiveness of the proposed defenses in the tested settings without establishing universal empirical superiority. The empirical ordering depends on the attack setting and may also depend on the feature map, attack budget, and regularization strength. The feature map

Method	No-attack	Random (30%)	Top-impact	Targeted BMP-A
DPO (no defense)	0 (ref.)	0.041	0.155	0.289
DPO_REG	0.069	0.087	0.109	0.130
TRIM	0	0.039	0.018	0.335
TRIM_REG	0.069	0.092	0.080	0.118
SOFT_TRIM	0	0.037	0.064	0.320
SOFT_TRIM_REG	0.069	0.089	0.093	0.123
DR_DPO (Wu et al., 2024b)	≈0	0.082	0.423	1.066
NR_DPO (Chowdhury et al., 2024)	0	0.010	0.304	1.079
WDPO (Xu et al., 2025)	0.096	0.098	0.101	0.103
KLDPO (Xu et al., 2025)	0.089	0.096	0.102	0.107

Table 1: Final-step candKL across all four baseline-comparison scenarios. The random-corruption column reports ten-seed means for all methods. Appendix I reports the corresponding mean and standard deviation for the proposed methods. In the no-attack control, the unregularized proposed defenses use $K = 0$ and the regularized ones keep $\lambda = 10^{-3}$. NR_DPO uses $\epsilon = 0$ without attacks and $\epsilon = 0.3$ in attacked scenarios.

determines impact magnitudes and gradient directions. Changing it may change the attacked/trimmed-set overlap and the relative performance of the defenses. The differences across attacks, together with the clean-data and random-flip controls, illustrate why our analysis separates clean defense bias from surviving attack effects. Building on the constant flip-gradient shift in Lemma 2.2, the resulting gradient-deviation bounds yield time-uniform parameter stability and then policy-level KL guarantees under the stated log-linear, bounded-region, and curvature assumptions. These guarantees hold uniformly over all fixed attack sets containing at most K flips, including top-impact attack, random attack, targeted attack, or even attacks with knowledge of the defenses.

K Experiments: Matched Regularization Controls

We compare TRIM_REG and SOFT_TRIM_REG with full-data DPO_REG on the same attacked data. All three methods use the quadratic penalty $\lambda\|\theta - \theta_{\text{ref}}\|_2^2/2$ with $\lambda = 10^{-3}$. DPO_REG does not trim or downweight pairs, allowing us to separate the effect of trimming from the shared quadratic regularizer and measure the additional effect of the impact-score rule.

Table 2: Final SGD candidate-set KL from full-data clean DPO in the 160,000-pair SHP setting. Percentages are computed from unrounded values relative to matched full-data DPO_REG.

Attack	DPO_REG	TRIM_REG	Reduction	SOFT_TRIM_REG	Reduction
Top-impact	0.1089	0.0803	26.3%	0.0934	14.2%
BMP-A	0.1296	0.1176	9.3%	0.1232	4.9%

Table 2 shows lower candKL for both proposed REG variants than for DPO_REG under both attacks. The reductions are larger for top-impact than for BMP-A. Figure 1b shows the corresponding BMP-A trajectories. Together, these comparisons show that impact-based trimming and downweighting provide an additional benefit beyond the shared quadratic regularizer in these two attack settings.

L Experiments: Score Assignment and BMP-A Magnitude Geometry

To test the effect of score assignment, Random TRIM_REG removes the same number of pairs as TRIM_REG, while Permuted SOFT_TRIM_REG randomly reassigns the SOFT_TRIM_REG weights, preserving their multiset and normalization. These are separate randomized controls, not aliases for the impact-based methods. Both controls randomize only among pairs with nonzero flip impact. The

35,256 zero-impact pairs remain untrimmed and have weight one. Within each paired comparison, the attacked set, candidate set, clean-DPO reference, K, and λ are fixed.

We use the SGD setup of Appendix G: 100,000 updates, learning rate 0.25, batch size 25,600, and candKL evaluation every 100 updates. We reuse the original SGD trajectories for `DPO_REG`, `TRIM_REG`, and `SOFT_TRIM_REG`. Hard-trim methods sample uniformly with replacement from their retained subsets. Soft-trim methods sample all pairs and normalize by the global mean weight. Each proposed method and its randomized control use the same initialization and minibatch seed.

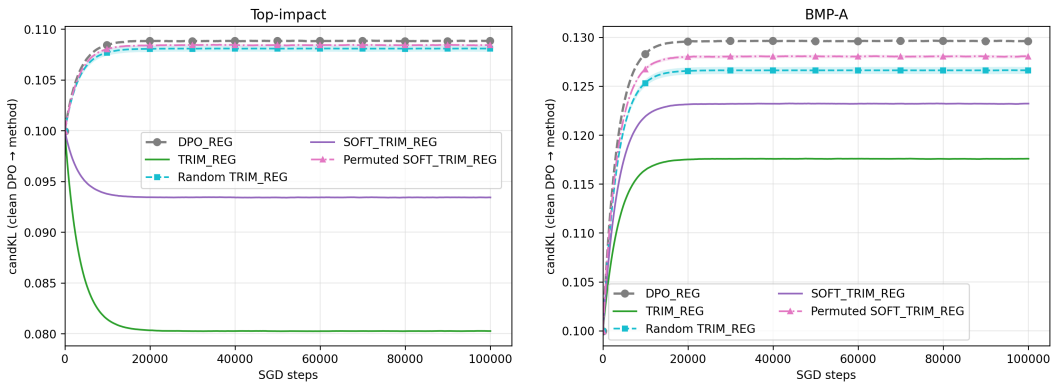
Table 3: Nonzero-conditioned control candKL minus impact-score candKL after 100,000 SGD updates. Positive values favor the impact score. Mean $\pm$ sample standard deviation is over ten paired trials.

Scenario	`TRIM_REG` vs. Random `TRIM_REG`	`SOFT_TRIM_REG` vs. Permuted `SOFT_TRIM_REG`
Clean, wrong K	-0.0086 ± 0.0002	-0.0042 ± 0.0001
Random flips	-0.0038 ± 0.0005	-0.0019 ± 0.0004
Top-impact	$+0.0278 \pm 0.0003$	$+0.0150 \pm 0.0001$
BMP-A	$+0.0090 \pm 0.0004$	$+0.0048 \pm 0.0002$

We repeat each comparison for ten randomized assignments, holding the minibatch seed fixed. Random-flip trials also use different attacks. Under top-impact and BMP-A, both impact-based methods have lower candKL than their randomized controls in all ten trials. The randomized controls have lower candKL in the clean-data and random-flip settings.

For BMP-A, the top-30% impact region contains 20,232 of the 48,000 flipped pairs (42.15%). These pairs account for 58.88% of the summed impact and 72.76% of the summed squared impact over the attacked set.

Figure 12 shows the candKL for the top-impact and BMP-A attacks. The gains over matched randomized controls show that the impact-score assignment matters for these attacks, beyond the number of retained pairs or the distribution of weights. In contrast, random flips do not preferentially affect high-impact pairs, and the clean-data control contains no attacked pairs to remove. The mixed results in Table 3 actually reflect the distinction between our uniform perturbation bounds and a claim that impact-based selection is best for every corruption pattern.



(a) Top-impact attack.

(b) BMP-A attack.

Figure 12: Score-assignment ablation: candKL versus SGD steps, with the same quadratic regularizer for every method. Random `TRIM_REG` and Permuted `SOFT_TRIM_REG` are the nonzero-conditioned controls. Lines show means and shaded bands show $\pm$ one sample standard deviation over ten random assignments with fixed minibatch seeds. The impact-based methods and `DPO_REG` use one fixed SGD run for each attack.

M EXPERIMENTS: THREE-DATASET BMP-A BUDGET SWEEP

We evaluate SHP (Ethayarajh et al., 2022), HH-RLHF (Bai et al., 2022; Ganguli et al., 2022), and UltraFeedback-Binarized (Cui et al., 2023). Each dataset uses $6{,}000$ training and $1{,}500$ held-out pairs for seeds $0, 1, 2$.

Data and features. SHP reuses the cached features and seed-0 $80/20$ split from Appendix G. HH-RLHF uses `Anthropic/hh-rlhf`, with its `train` and `test` splits. UltraFeedback-Binarized uses `HuggingFaceH4/ultrafeedback_binarized`, with its `train_prefs` and `test_prefs` splits. We retain the last human prompt and final assistant response for HH-RLHF, and the prompt and final assistant response for UltraFeedback-Binarized. Empty or identical response pairs are excluded, as are UltraFeedback pairs with tied preference scores. We construct pools of $10{,}000$ training and $1{,}600$ held-out pairs per external dataset, excluding held-out prompts that occur in the sampled training pool, and subsample the experimental pairs independently for each seed.

External-dataset features are the ℓ_2-normalized 384-dimensional `[CLS]` outputs of frozen `all-MiniLM-L6-v2`. The input concatenates the last 800 prompt characters with the first $5{,}000$ response characters, separated by `Assistant:`, and is truncated to 256 tokens. Feature differences are not normalized, preserving the variation in flip-impact scores.

Optimization and attack. All methods start at $\theta_0 = \theta_{\mathrm{ref}} = 0$ and use $\beta = 1$, $100{,}000$ SGD steps, batch size $1{,}024$, learning rate 0.25, $\lambda = 10^{-3}$ for REG, and $\tau = 5K/12$. Seeds vary data subsampling and SGD. Within each seed, methods use the same data, attack, and minibatch seed. Following Appendix H, the target is the reflection $\theta^\dagger = 2\theta_{\mathrm{ref}} - \theta^\star = -\theta^\star$ of the final clean-DPO parameter. We run BMP-A once to the largest budget and use nested prefixes for the $5\%, 10\%, 20\%, 30\%$ attacks. Each attacked set remains fixed during training.

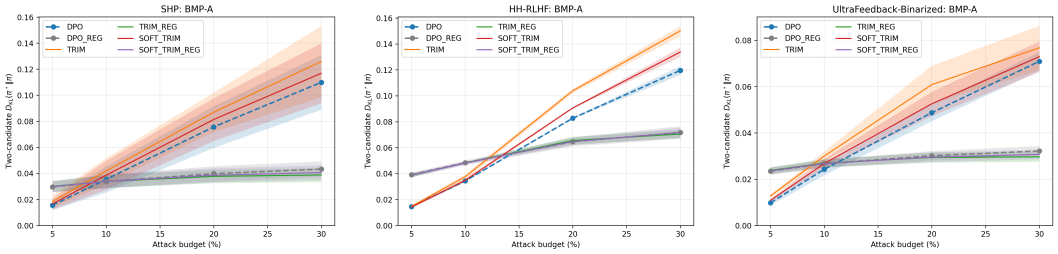
Metric. Table 4 and Figure 13 report held-out two-candidate $D_{\mathrm{KL}}(\pi_{\mathrm{clean}} \| \pi_{\mathrm{method}})$, with each candidate set consisting of the held-out preferred and rejected responses. This metric is not directly comparable to the main paper's 100-candidate candKL.

Table 4: Three-seed mean two-candidate policy KL under nested BMP-A budgets. Residual is BMP-A's relative target-gradient residual.

Dataset	Budget	DPO	DPO_REG	TRIM	TRIM_REG	SOFT_TRIM	SOFT_TRIM_REG	BMP-A Residual
SHP	5%	.015	.030	.018	.030	.017	.030	.674
	10%	.035	.034	.042	.034	.039	.034	.493
	20%	.076	.040	.087	.038	.081	.039	.244
	30%	.110	.043	.126	.039	.117	.041	.102
HH-RLHF	5%	.015	.039	.015	.039	.014	.039	.698
	10%	.034	.048	.038	.048	.035	.048	.520
	20%	.083	.065	.104	.066	.091	.065	.271
	30%	.120	.072	.150	.071	.134	.072	.174
UltraFeedback-Binarized	5%	.010	.023	.013	.024	.011	.024	.675
	10%	.024	.027	.030	.027	.027	.027	.478
	20%	.049	.030	.061	.029	.053	.030	.273
	30%	.071	.032	.077	.030	.073	.031	.151

At 20% and 30%, both proposed REG variants have lower KL than DPO on all three datasets. Their differences from `DPO_REG` are smaller, and both have slightly higher KL than `DPO_REG` on HH-RLHF at 20% before rounding. At 5%, all REG methods have higher KL than DPO. The smaller-budget attacks also have larger relative target-gradient residuals, ranging from 0.674 to 0.698 at 5% and from 0.478 to 0.520 at 10%. These residuals indicate that BMP-A less closely matches the target gradient at smaller budgets, and attacked DPO remains closer to the clean policy than at the larger budgets.

Overall, the regularized defenses are effective against the larger-budget BMP-A attacks across all three datasets. At smaller budgets, defense-induced bias can outweigh the reduction in the surviving-attack effect. This is the tradeoff captured by our bounds, which do not assert that a defense improves on DPO for every attacked set.



(a) SHP. (b) HH-RLHF. (c) UltraFeedback-Binarized.

Figure 13: Three-dataset log-linear BMP-A budget sweep. Lines show means and shaded bands show $\pm$ one sample standard deviation over three seeds.

N Experiments: Budget Selection and Misspecified K

Our bounds assume that K is an upper bound on the number of flips. Here we examine sensitivity to the assumed budget and select K using held-out policy KL to the clean-DPO reference available in the experiment.

We use the datasets, preprocessing, initialization, optimizer, and reflected-target construction of Appendix M. For each dataset and seed, one fixed 20% true BMP-A attacked set on $6{,}000$ training pairs is reused while the defender's assumed budget varies over $K/N \in \{0\%, 1\%, 2\%, 4\%, 8\%, 16\%, 32\%\}$. For each method and seed, we select the budget with the lowest two-candidate KL on the first 500 held-out pairs, breaking ties toward smaller K. We evaluate the selected method on the remaining $1{,}000$ pairs. There are three seeds, $100{,}000$ SGD steps, $\lambda = 10^{-3}$, and $\tau = 5K/12$ for $K > 0$. When $K = 0$, all soft weights are set to one, so the unregularized and REG methods reduce to DPO and `DPO_REG`, respectively.

Table 5: Mean test two-candidate KL for a fixed true 20% attack while assumed K changes. Each entry is unregularized / regularized methods.

Dataset	Assumed K/N	`TRIM` / `TRIM_REG`	`SOFT_TRIM` / `SOFT_TRIM_REG`
SHP	0%	.0743 / .0390	.0743 / .0390
	1%	.0767 / .0390	.0758 / .0390
	2%	.0770 / .0388	.0760 / .0389
	4%	.0788 / .0386	.0765 / .0388
	8%	.0806 / .0381	.0780 / .0385
	16%	.0875 / .0375	.0801 / .0381
	32%	.0886 / .0357	.0820 / .0372
HH-RLHF	0%	.0818 / .0636	.0818 / .0636
	1%	.0824 / .0640	.0824 / .0640
	2%	.0820 / .0636	.0821 / .0639
	4%	.0832 / .0632	.0819 / .0635
	8%	.0872 / .0634	.0837 / .0634
	16%	.0983 / .0643	.0877 / .0637
	32%	.1099 / .0628	.0941 / .0639
UltraFeedback-Binarized	0%	.0484 / .0303	.0484 / .0303
	1%	.0500 / .0303	.0493 / .0302
	2%	.0498 / .0301	.0494 / .0302
	4%	.0508 / .0299	.0497 / .0301
	8%	.0521 / .0297	.0502 / .0299
	16%	.0557 / .0294	.0512 / .0297
	32%	.0595 / .0288	.0531 / .0294

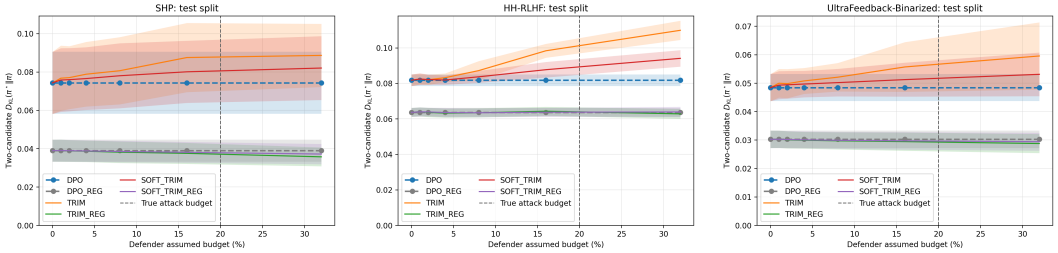
Figure 14 shows the held-out test curves as the assumed budget varies. Both proposed REG variants have lower mean test KL than DPO across the entire grid on all three datasets, while unregularized

Table 6: Validation-selected assumed K/N for seeds $0/1/2$. The true attack fraction is 20% in every run.

Dataset	TRIM	TRIM_REG	SOFT_TRIM	SOFT_TRIM_REG
SHP	0/0/0	32/32/32	0/0/0	32/32/32
HH-RLHF	0/0/0	32/32/4	0/0/0	32/16/4
UltraFeedback-Binarized	2/0/0	32/32/32	8/0/0	32/32/32

trimming generally increases the deviation. Validation selects the test-grid minimizer in 31 of 36 cases. Only 15 of the 36 selected budgets satisfy $\widehat{K}/N \geq 20\%$.

Thus, the REG variants retain their advantage over attacked DPO in this experiment even when the assumed budget differs from the true budget. A clean-policy reference also helps select K for policy proximity, but the selected budget is not an estimate of the number of attacked pairs. The uniform guarantees apply when the chosen K is an upper bound on that number.



(a) SHP. (b) HH-RLHF. (c) UltraFeedback-Binarized.

Figure 14: Held-out test KL for a fixed-20% BMP-A support with a varied defender budget. Lines show means and shaded bands show $\pm$ one sample standard deviation over three seeds. The vertical dashed line marks the true attack fraction.

O Extension to Nonlinear DPO Policies

We discuss the difficulties in extending the analysis beyond log-linear policies and report preliminary experiments with a small nonlinear policy over frozen representations.

O.1 Flip-gradient identity for differentiable DPO policies

We restate the flip-gradient identity of Yang et al. (2026) in our notation and include its proof for completeness.

Lemma O.1 (Flip-gradient identity for differentiable DPO policies)**.** *For any differentiable policy π_θ and a flipped pair $i \in A$, let ℓ_i be the DPO loss on the clean pair (x^i, y_w^i, y_l^i) and let $\tilde{\ell}_i$ be the observed loss after swapping the two responses. Wherever the log-policy gradients exist,*

$$\nabla \tilde{\ell}_i(\theta) - \nabla \ell_i(\theta) = \beta \left[\nabla \log \pi_\theta(y_w^i \mid x^i) - \nabla \log \pi_\theta(y_l^i \mid x^i) \right]. \tag{22}$$

At a fixed θ, reversing the pair orientation negates the vector on the right-hand side and therefore preserves its norm. Unlike Lemma 2.2, however, this vector generally depends on θ.

Proof. Let $z_i(\theta)$ be the DPO logit of the clean pair, so that $\ell_i(\theta) = -\log \sigma(z_i(\theta))$ and $\tilde{\ell}_i(\theta) = -\log \sigma(-z_i(\theta))$. Differentiation gives $\nabla \tilde{\ell}_i(\theta) - \nabla \ell_i(\theta) = \nabla z_i(\theta)$. The reference-policy terms in $z_i(\theta)$ are constant in θ, leaving exactly (22). □

O.2 Why the log-linear robustness analysis does not directly extend

In the log-linear class, each flip contributes the fixed gradient shift βv_i, so the impact ranking and attack-induced gradient shift do not change during training. For a nonlinear policy, the flip-gradient

vector in Lemma O.1 depends on θ. Extending our analysis would require control of these vectors and their variation over the training region, rather than only at initialization. Recomputing scores can also change the retained set or weights during training.

The nonlinear objective is generally nonconvex, so the covariance argument used in Theorems 3.2 and 3.5 does not directly provide the required curvature and contraction. A nonlinear extension would also need a suitable parameter-to-policy bound and, for refreshed scores, an analysis of the changing objective.

O.3 Nonlinear preliminary experiments setup

We use SHP with seeds $0, 1, 2$, subsampling $6{,}000$ training pairs and $2{,}000$ held-out pairs from the split in Appendix G. A seed-specific Gaussian projection followed by ℓ_2 normalization maps frozen 384-dimensional MiniLM representations to $\bar{\psi} \in \mathbb{R}^{32}$. The policy uses the nonlinear energy head

$$f_\theta(\bar{\psi}) = a^\top \tanh(W\bar{\psi} + b), \qquad \pi_\theta(j \mid x; \mathcal{Y}) \propto \exp\big(f_\theta(\bar{\psi}(x, \mathcal{Y}_j(x)))\big), \tag{23}$$

with hidden width 8. Seeds vary data subsampling, the projection, initialization, and SGD. Within each seed, all methods share the same randomly initialized frozen reference, starting parameters, and minibatch seed. We run $10{,}000$ SGD steps with batch size 512, learning rate 1, $\beta = 1$, $\lambda = 10^{-3}$ for REG, and $\tau = 5K/12$.

The attack target is trained from the reference for $10{,}000$ steps with all preference labels reversed. Neural BMP-A uses the exact flip-gradient vectors at that target and constructs nested attack sets for $5\%, 10\%, 20\%, 30\%$ budgets. Each set is fixed during the corresponding training run. The defender does not observe the target or attacked set and computes the static impact scores once at the common reference initialization. At 30% ($K = 1{,}800$), the mean BMP-A relative target residual is 0.097, and the static top-K set contains 40.9% of the attacked pairs. We also train on clean data using the same assumed budgets to measure the clean defense bias.

The $2{,}000$ held-out pairs are used for pairwise diagnostics. The reported candKL instead uses 500 fixed held-out SHP prompts from the candidate cache of Appendix G, with 100 candidates per prompt, shared across methods. The clean-DPO nonlinear policy trained for the same number of steps is the comparator.

O.4 Static impact-score results

Table 7: Mean $\pm$ sample standard deviation candidate-set $D_{\mathrm{KL}}(\pi_{\mathrm{clean}} \| \pi_{\mathrm{method}})$ over three nonlinear-head seeds under the fixed 30% neural BMP-A setting.

DPO	DPO_REG	TRIM	TRIM_REG	SOFT_TRIM	SOFT_TRIM_REG
$.213 \pm .025$	$.141 \pm .019$	$.297 \pm .029$	$.115 \pm .013$	$.241 \pm .009$	$.128 \pm .015$

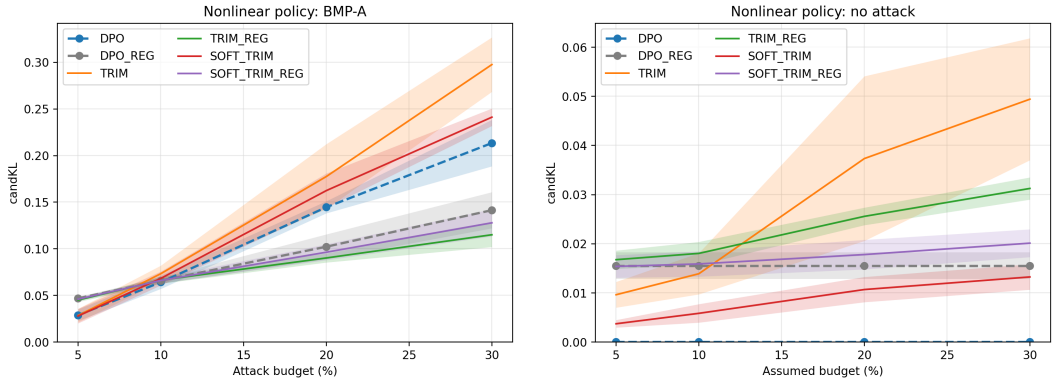
At the 30% attack budget, TRIM_REG and SOFT_TRIM_REG have 18.7% and 9.6% lower mean candKL than DPO_REG, respectively (Table 7). Without regularization, TRIM and SOFT_TRIM have higher candKL than DPO.

Figure 15 shows the budget sweep and clean-data controls. At 5%, all REG methods, including full-data DPO_REG, have higher candKL than DPO, while unregularized TRIM and SOFT_TRIM slightly improve on DPO. On clean data, candKL increases with the assumed budget for unregularized TRIM and SOFT_TRIM. At 30%, their candKL values are 0.0494 and 0.0132, respectively.

These preliminary results show a benefit from combining impact-based trimming with regularization under the larger nonlinear BMP-A attack. At smaller budgets, the regularization bias can outweigh this benefit, as also seen for full-data DPO_REG. Extending the robustness guarantees to this policy class requires the additional analysis discussed in Appendix O.2.

O.5 Periodic impact-score refresh

We keep the 30% attacked set fixed and recompute the parameter-dependent impact scores every 500 optimization steps. This gives 20 score evaluations at steps $0, 500, \ldots, 9{,}500$.



(a) Neural BMP-A attacks.

(b) Clean data, assumed attack budgets.

Figure 15: Small nonlinear-head sweep over frozen SHP representations. Left: nested $5\%, 10\%, 20\%, 30\%$ neural BMP-A attacks. Right: clean data with the same assumed budgets. Lines show means and shaded bands show $\pm$ one sample standard deviation over three seeds.

Table 8: Static versus periodically refreshed candidate-set KL, averaged over three nonlinear-head seeds.

Method	Static	Refreshed
TRIM	.2974	.2941
TRIM_REG	.1149	.1056
SOFT_TRIM	.2410	.2266
SOFT_TRIM_REG	.1277	.1216

Table 8 shows lower mean candKL for all four methods after score refresh. At the last refresh, the top-K set contains 43.6% of the attacked pairs for TRIM and 45.8% for SOFT_TRIM, compared with 40.9% at initialization. The corresponding fractions for TRIM_REG and SOFT_TRIM_REG are 41.5% and 41.9%. TRIM removes these pairs, while SOFT_TRIM downweights them.

Refreshing the impact scores therefore gives a further improvement in this nonlinear experiment, with TRIM_REG attaining the lowest mean candKL among the compared methods. The unregularized variants still have higher candKL than DPO, so score refresh alone does not resolve the tradeoff between removing attacked pairs and changing the clean objective.