TPO: Tri-level Distributionally Robust Learning for OOD Direct Preference Optimization

Chengtao Jian[1], **Kai Yang**[1,2,3*]
[1]School of Computer Science and Technology, Tongji University, Shanghai, China
[2]Shenzhen Loop Area Institute, Shenzhen, China
[3]Shanghai Research Institute for Intelligent Autonomous Systems, Shanghai, China
{jct,kaiyang}@tongji.edu.cn

Abstract

Direct Preference Optimization (DPO) minimizes empirical risk over observed preference comparisons, implicitly assuming that the training preference distribution matches the deployment preference distribution. In practice, this distributional assumption often fails, as preference distribution shifts can arise at three levels: group-level shift, prompt-level shift, and preference-level shift. Existing robust DPO methods typically either model uncertainty over the joint preference distribution or focus on individual sources of shift, leaving their hierarchical structure underexplored. To address this gap, we propose TPO, a tri-level distributionally robust framework for out-of-distribution (OOD) preference alignment, which jointly models all three levels of distributional uncertainty through a hierarchically structured rectangular ambiguity set. By instantiating the ambiguity set with Wasserstein distances, we derive a tractable and interpretable objective that combines a preference-level margin penalty, prompt-level adversarial embedding perturbation, and worst-case group reweighting, each governed by an independent radius. We establish finite-sample convergence guarantees and empirically validate the effectiveness of TPO across diverse OOD environments.

1 Introduction

Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as a prominent method for aligning large language models (LLMs) with human preferences. Unlike RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022), DPO directly optimizes a preference objective from pairwise comparison data, bypassing explicit reward modeling and improving optimization stability. However, standard DPO minimizes empirical risk over observed pairwise preference data, without explicitly accounting for distribution shifts between training and deployment. Such out-of-distribution (OOD) shifts can degrade preference alignment performance (Wu et al., 2025; Xu et al., 2025).

In preference alignment, OOD preference shift exhibits a natural hierarchical structure across multiple levels. **First**, the distribution over groups may change between training and deployment, inducing a *group-level shift*, which has been studied in group robust preference alignment (Ramesh et al., 2024; Sagawa et al., 2020). **Second**, even within the same group, the prompt marginal distribution may vary across domains, time, or user interaction patterns, yielding a *prompt-level shift*, which corresponds to covariate shift in the prompt distribution and has been observed in robust prompt optimization (Li et al., 2023; Seegmiller and Preum, 2026). **Third**, the conditional preference distribution may also change due to noisy feedback or heterogeneous preferences within the same group, giving rise to a *preference-level shift* (Yao et al., 2025; Chowdhury et al., 2024; Poddar et al., 2024).

*Corresponding author.

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

Table 1: Comparison of robust DPO methods by the shift levels explicitly modeled in their objectives. ✓ denotes a dedicated robustness component, while ✗ denotes no explicit level-specific component.

Methods	Overall	Group-level	Prompt-level	Preference-level
DPO (Rafailov et al., 2023)	✗	✗	✗	✗
WDPO / KLDPO (Xu et al., 2025)	✓	✗	✗	✗
GRPO (Ramesh et al., 2024)	✗	✓	✗	✗
rDPO (Chowdhury et al., 2024)	✗	✗	✗	✓
Dr.DPO (Wu et al., 2025)	✗	✗	✗	✓
TPO (Ours)	✓	✓	✓	✓

Recent robust DPO methods have begun to address distributional robustness, but they typically do not explicitly model the hierarchical structure of preference distribution shift. WDPO and KLDPO (Xu et al., 2025) construct Wasserstein and KL ambiguity sets over the entire preference distribution, while rDPO (Chowdhury et al., 2024) and Dr.DPO (Wu et al., 2025) focus on preference label noise and on pointwise and pairwise data noise, respectively. As summarized in Table 1, existing methods either model overall uncertainty or focus on a single robustness level. Motivated by these observations, we propose TPO, a tri-level distributionally robust framework for OOD preference alignment, which jointly models all three levels through a hierarchically structured ambiguity set. Overall, our main contributions are as follows:

- **Tri-level learning for OOD preference alignment.** We identify that preference distribution shift exhibits a natural tri-level structure comprising group-level, prompt-level, and preference-level shifts that existing robust DPO methods do not jointly address. We propose TPO, which captures all three levels within a hierarchical rectangular ambiguity set. This decomposition is relevant to safe LLM alignment, as real-world LLMs inevitably encounter distributional mismatch at multiple levels simultaneously, while single-level robustness leaves models vulnerable to the remaining sources of shift.
- **Tractable Wasserstein robust preference objective.** By instantiating the hierarchical ambiguity set with Wasserstein distances, we derive a tractable and interpretable training objective with three complementary components: a preference-level margin penalty, prompt-level adversarial embedding perturbation, and worst-case group reweighting. Each component targets a distinct source of distributional uncertainty and is governed by its own radius, enabling flexible and modular control over robustness.
- **Theoretical guarantees and empirical validation.** We establish finite-sample convergence guarantees, showing that the empirical minimizer converges to the population minimizer at a rate of $\widetilde{\mathcal{O}}(n_{\min}^{-1/4})$, where $n_{\min} = \min_{g\in[m]} n_g$ is the smallest group size. Empirically, TPO achieves superior or competitive performance over standard DPO and robust DPO baselines across multiple OOD shift settings.

2 Preliminaries

Notation. Unless otherwise specified, we use uppercase letters for random variables and lowercase letters for their realizations, with probability distributions denoted by P and Q. We use $\boldsymbol{x} \in \mathcal{X}$ for a prompt, $\boldsymbol{y} \in \mathcal{Y}$ for a response, and $\boldsymbol{h} \in \mathcal{H}$ for the prompt embedding, with $\tilde{\boldsymbol{h}} = \boldsymbol{h} + \boldsymbol{\delta}$ denoting an adversarially perturbed embedding. A pairwise preference example is modeled as a random variable $U = (X, V, S)$, where $V = (Y_1, Y_2)$ is a pair of candidate responses and $S \in \{+1, -1\}$ is the binary preference label indicating which response is preferred. In standard DPO training, each comparison $(\boldsymbol{x}, \boldsymbol{y}^+, \boldsymbol{y}^-)$ is encoded as $V = (\boldsymbol{y}^+, \boldsymbol{y}^-)$ and $S = +1$. The group index is $g \in \mathcal{G} = [m]$, where $[m] = \{1, \dots, m\}$, and P_g denotes the group-specific distribution over U. Let I_g be the index set of the $n_g = |I_g|$ training samples from group g. We factorize $P_g(U) = P_g^X(X)\, P_g^V(V \mid X)\, P_g^S(S \mid V, X)$, where P_g^X, P_g^V, and P_g^S are the prompt marginal distribution, conditional candidate response distribution, and conditional preference distribution, respectively. We write Δ^{m-1} for the probability simplex over m groups, $\|\boldsymbol{v}\|_2$ for the Euclidean norm, $\phi(\boldsymbol{h}, \boldsymbol{y}) \in \mathbb{R}^d$ for the log-linear feature map, and $\varphi(t) = \log(1 + e^{-t})$ for the logistic loss.

Direct Preference Optimization. DPO learns a policy $\pi_{\boldsymbol{\theta}}$ from pairwise preference data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-)\}_{i=1}^n$, where $\boldsymbol{y}_i^+ \succ \boldsymbol{y}_i^-$ indicates that $\boldsymbol{y}_i^+$ is preferred over $\boldsymbol{y}_i^-$. The DPO margin for a

candidate pair $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ is defined as

$$m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) = \beta\left(\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_1 \mid \boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}_1 \mid \boldsymbol{x})} - \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_2 \mid \boldsymbol{x})}{\pi_{\mathrm{ref}}(\boldsymbol{y}_2 \mid \boldsymbol{x})}\right), \tag{1}$$

where $\beta > 0$ is a scaling parameter and π_{ref} is a reference policy. Given a binary preference label $s \in \{+1, -1\}$, the labeled DPO loss is $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2, s; \boldsymbol{\theta}) = \varphi(s \cdot m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}))$. For the standard DPO setting, we set $(\boldsymbol{y}_1, \boldsymbol{y}_2) = (\boldsymbol{y}^+, \boldsymbol{y}^-)$ and $s = +1$, which gives the empirical DPO risk minimization objective:

$$\widehat{\mathcal{R}}_{\mathrm{DPO}}(\boldsymbol{\theta}) := \mathbb{E}_{\hat{P}}\big[\mathcal{L}(U; \boldsymbol{\theta})\big] = \frac{1}{n}\sum\nolimits_{i=1}^{n} \varphi\big(m(\boldsymbol{x}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-; \boldsymbol{\theta})\big). \tag{2}$$

Standard DPO minimizes this empirical risk under the assumption that training and deployment preferences follow the same distribution. When this distributional assumption fails, DPO can suffer from reduced preference alignment quality (Wu et al., 2025; Xu et al., 2025).

Group DRO. Group DRO (Sagawa et al., 2020) introduces a group variable $G \in \mathcal{G} = \{1, \ldots, m\}$, so that each preference observation $U = (X, V, S)$ is associated with a group index g. Let P_g denote the population distribution over U for group g. The overall training distribution is expressed as $P = \sum_{g=1}^{m} \alpha_g P_g$, where $\boldsymbol{\alpha} \in \Delta^{m-1}$ are the group proportions. Through the lens of distributionally robust optimization (Jiao et al., 2022; Esfahani and Kuhn, 2015), Group DRO replaces the ERM objective in Eq. (2) with the following min-max optimization problem:

$$\inf_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{\omega} \in \mathcal{Q}} \sum\nolimits_{g=1}^{m} \omega_g \, \mathbb{E}_{P_g}\big[\mathcal{L}(U; \boldsymbol{\theta})\big], \tag{3}$$

which minimizes the worst-case expected loss over an ambiguity set $\mathcal{Q}$ of group mixture weights, a paradigm that has also been adopted in robust RLHF (Chakraborty et al., 2024; Ramesh et al., 2024). However, this formulation keeps the within-group distributions P_g fixed and does not explicitly model shifts in either the prompt marginal distribution (Li et al., 2023; Koh et al., 2021) or the conditional preference distribution (Sorensen et al., 2024; Yao et al., 2025). This motivates the hierarchical robust formulation in Section 3.1.

3 Method

3.1 Hierarchical Distribution Shifts and Uncertainty in Preference Alignment

In preference alignment, distribution shift exhibits a natural three-level hierarchical structure. At the *global* level, the group proportions may shift between training and deployment (Qiao et al., 2025; Ramesh et al., 2024; Jian et al., 2024). Within each group, shift arises at the *local* level in two forms: the prompt marginal distribution P_g^X may shift across environments (Li et al., 2023; Koh et al., 2021), and the conditional preference distribution $P_g^S(S \mid V, X)$ may vary due to noisy (Chowdhury et al., 2024) or pluralistic (Poddar et al., 2024; Yao et al., 2025) human feedback. We refer to these as *group-level*, *prompt-level*, and *preference-level* distribution shifts, respectively. We next formalize these three levels of distributional uncertainty and construct a hierarchical ambiguity set for robust preference optimization.

Global group ambiguity set. At the global level, motivated by group DRO (Sagawa et al., 2020), we model group-level shift by an ambiguity set over the group mixture weights, centered at the training group proportions $\boldsymbol{\alpha}$: $\mathcal{U}^{\omega} = \big\{\boldsymbol{\omega} \in \Delta^{m-1} : \mathcal{D}_{\omega}(\boldsymbol{\omega}, \boldsymbol{\alpha}) \le \rho^{\omega}\big\}$, where $\mathcal{D}_{\omega}$ is a discrepancy measure on the group simplex and $\rho^{\omega} \ge 0$ is the group-level radius.

Local rectangular ambiguity sets. At the local level, within-group shift arises from two sources: the prompt marginal distribution and the conditional preference distribution. For each group g, the shifted joint distribution factorizes as $Q_g(U) = Q_g^X(X)\, P_g^V(V \mid X)\, Q_g^S(S \mid X, V)$, where Q_g^X and Q_g^S are each constrained by their own ambiguity set. Since the candidate response pairs are fixed in the offline preference dataset (Rafailov et al., 2023), we do not introduce an ambiguity set for P_g^V and keep $P_g^V(V \mid X)$ unchanged. We first constrain the prompt marginal distribution by $\mathcal{U}_g^X = \{Q_g^X : \mathcal{D}_X(Q_g^X, P_g^X) \le \rho_g^X\}$, where $\mathcal{D}_X$ is a discrepancy measure on the prompt (embedding) space and $\rho_g^X \ge 0$ is the prompt-level radius.

For each prompt and response pair $(\boldsymbol{x}, \boldsymbol{v})$, we constrain the conditional preference distribution by

$$\mathcal{U}^S_{g,\boldsymbol{x},\boldsymbol{v}} = \Big\{ Q^S_g(\cdot \mid \boldsymbol{x}, \boldsymbol{v}) : \mathcal{D}_S\big(Q^S_g(\cdot \mid \boldsymbol{x}, \boldsymbol{v}), P^S_g(\cdot \mid \boldsymbol{x}, \boldsymbol{v})\big) \le \rho^S_g \Big\}, \tag{4}$$

where $\mathcal{D}_S$ measures discrepancies between distributions over binary preference labels, and $\rho^S_g \in [0, 1]$ is the preference-level radius. Jointly, these constraints induce the local ambiguity set

$$\mathcal{U}_g = \Big\{ Q_g : Q^X_g \in \mathcal{U}^X_g,\ Q^S_g(\cdot \mid \boldsymbol{x}, \boldsymbol{v}) \in \mathcal{U}^S_{g,\boldsymbol{x},\boldsymbol{v}} \text{ for } Q^X_g P^V_g\text{-a.e. } (\boldsymbol{x}, \boldsymbol{v}) \Big\}. \tag{5}$$

Note that $\mathcal{U}_g$ is *rectangular* with respect to this factorization: the prompt marginal and the conditional preference distribution can be chosen separately under their respective constraints. Under standard measurability conditions, this rectangularity allows the adversary to first select a worst-case preference distribution for each fixed (X, V), and then a worst-case prompt marginal distribution (Shapiro, 2016; Li and Shapiro, 2025), which yields the nested worst-case optimization in Section 4.2.

Hierarchical ambiguity set. Inspired by structured uncertainty modeling (Jo et al., 2026; Yang et al., 2014), we combine the global and local constraints to construct the full ambiguity set:

$$\mathcal{U} = \Big\{ Q = \textstyle\sum_{g=1}^m \omega_g Q_g : \boldsymbol{\omega} \in \mathcal{U}^\omega,\ Q_g \in \mathcal{U}_g \text{ for all } g \in \mathcal{G} \Big\}. \tag{6}$$

Compared with group robustness methods that only consider group-level shift, our formulation more faithfully characterizes preference distribution shift by explicitly accounting for within-group variation in both the prompt marginal and the conditional preference distribution.

3.2 Tri-level Distributionally Robust Preference Optimization

Based on the hierarchical ambiguity set $\mathcal{U}$ in Eq. (6), we now formulate the robust preference optimization problem. Since the true distribution is unknown but assumed to lie in $\mathcal{U}$, we follow standard DRO principles (Wiesemann et al., 2014; Esfahani and Kuhn, 2015; Sinha et al., 2018) and minimize the worst-case risk over $\mathcal{U}$:

$$\inf_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum\nolimits_{g=1}^m \omega_g \sup_{Q_g \in \mathcal{U}_g} \mathbb{E}_{U \sim Q_g}\big[\mathcal{L}(U; \boldsymbol{\theta})\big]. \tag{7}$$

The outer maximization over $\boldsymbol{\omega}$ captures global group-level shift, and the inner maximization over Q_g captures local prompt- and preference-level shifts. To solve the inner maximization, we first define the worst-case preference loss of group g at a fixed prompt and response pair $(\boldsymbol{x}, \boldsymbol{v})$ as

$$\psi_g(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta}) := \sup_{Q^S_g \in \mathcal{U}^S_{g,\boldsymbol{x},\boldsymbol{v}}} \mathbb{E}_{S \sim Q^S_g}\big[\mathcal{L}(\boldsymbol{x}, \boldsymbol{v}, S; \boldsymbol{\theta})\big]. \tag{8}$$

Since $\mathcal{U}_g$ is rectangular, the adversarial choices of the prompt marginal Q^X_g and the conditional preference kernel $Q^S_g(\cdot \mid X, V)$ are decoupled, which enables a nested decomposition of the worst-case expectation over Q_g (Shapiro, 2016; Li and Shapiro, 2025). Consequently, Eq. (7) can be rewritten as

$$\inf_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum\nolimits_{g=1}^m \omega_g \sup_{Q^X_g \in \mathcal{U}^X_g} \mathbb{E}_{X \sim Q^X_g,\, V \sim P^V_g(\cdot|X)}\big[\psi_g(X, V; \boldsymbol{\theta})\big]. \tag{9}$$

The above nested optimization captures the three sources of distribution shift described in Section 3.1. It generalizes Group DRO by accounting for within-group distributional uncertainty: when $\rho^X_g = \rho^S_g = 0$ for all groups, the formulation reduces to a Group DRO objective over $\mathcal{U}^\omega$. While nested optimization structures have been studied in bilevel learning (Franceschi et al., 2018; Liu et al., 2025; Jiao et al., 2023; Ji et al., 2021) and trilevel learning (Jiao et al., 2025b; Jian et al., 2024; Yang et al., 2026), our formulation arises from hierarchical distributional uncertainty in preference alignment.

3.3 Finite-Sample Guarantees

We establish a finite-sample convergence guarantee for the fixed-penalty Wasserstein dual objective of TPO, whose derivation is presented in Section 4.2. The analysis relies on the following two assumptions.

Assumption 1 (Log-linear policy class (Xu et al., 2025) and boundedness (Jiao et al., 2024)) *Let $\phi : \mathcal{H} \times \mathcal{Y} \to \mathbb{R}^d$ be a feature map satisfying $\sup_{\boldsymbol{h}\in\mathcal{H},\, \boldsymbol{y}\in\mathcal{Y}} \|\phi(\boldsymbol{h}, \boldsymbol{y})\|_2 \le C_\phi$, where $\mathcal{H}$ includes both clean and perturbed prompt embeddings. Let $\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le B\}$, and $\pi_{\mathrm{ref}} = \pi_{\boldsymbol{\theta}_{\mathrm{ref}}}$ satisfy $\|\boldsymbol{\theta}_{\mathrm{ref}}\|_2 \le B$. We consider the log-linear policy class*

$$\Pi = \left\{ \pi_{\boldsymbol{\theta}} : \pi_{\boldsymbol{\theta}}(y \mid \boldsymbol{h}) = \frac{\exp(\boldsymbol{\theta}^\top \phi(\boldsymbol{h}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp(\boldsymbol{\theta}^\top \phi(\boldsymbol{h}, \boldsymbol{y}'))},\ \boldsymbol{\theta} \in \Theta \right\}.$$

Assumption 2 (Uniform covariance coverage (Zhu et al., 2023)) *Let $z(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-) = \phi(\boldsymbol{h}, \boldsymbol{y}^+) - \phi(\boldsymbol{h}, \boldsymbol{y}^-)$. For each group g and $\boldsymbol{\theta} \in \Theta$, assume the inner maximization in Eq. (20) admits a measurable maximizer $\boldsymbol{\delta}^\star_{g,\boldsymbol{\theta}}(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-)$, and let $\tilde{\boldsymbol{h}} = \boldsymbol{h} + \boldsymbol{\delta}^\star_{g,\boldsymbol{\theta}}(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-)$. There exists $\lambda > 0$ such that, for all $\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\omega} \in \mathcal{U}^\omega$, and all such maximizers, $\sum_{g=1}^m \omega_g \, \mathbb{E}_{P_g}\left[z(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-)\, z(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-)^\top \right] \succeq \lambda I_d$.*

For each group g, we fix $\gamma_g > 0$ in Eq. (19) and define the empirical risk as $\widehat{\mathcal{R}}_g(\boldsymbol{\theta}) = \gamma_g \rho_g^X + \frac{1}{n_g} \sum_{i \in I_g} \ell^g_{\boldsymbol{\theta}}(\boldsymbol{z}_i; \gamma_g)$, where $\ell^g_{\boldsymbol{\theta}}$ is given by Eq. (20). The corresponding population risk $\mathcal{R}_g(\boldsymbol{\theta})$ replaces the empirical average with $\mathbb{E}_{P_g}$. The hierarchical population and empirical objectives are $\mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\omega}\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \mathcal{R}_g(\boldsymbol{\theta})$ and $\widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\omega}\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \widehat{\mathcal{R}}_g(\boldsymbol{\theta})$, respectively. Let $\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\theta}} \in \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta})$ denote the population and empirical minimizers, respectively. The following theorem establishes a convergence rate governed by the smallest group sample size $n_{\min} = \min_{g\in[m]} n_g$.

Theorem 1 (Finite-sample convergence of TPO) *Let $M_\beta = 4\beta B C_\phi$ and $\kappa_\beta = \frac{\beta^2 e^{M_\beta}}{(1+e^{M_\beta})^2}$. Suppose Assumptions 1 and 2 hold. With probability at least $1-\delta$, we have*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \le 2\sqrt{\frac{\max_{g\in[m]} \left[2A_g \sqrt{\frac{d}{n_g}} + K_g \sqrt{\frac{\log(2m/\delta)}{2n_g}} \right]}{\kappa_\beta \lambda}},$$

where $K_g = \log(1 + e^{M_\beta}) + \rho_g^S M_\beta$, $L_g = 2\beta C_\phi (1 + \rho_g^S)$, $A_g = C K_g \sqrt{\log\big(e + 3BL_g/K_g\big)}$, and $C > 0$ is a universal constant. The proof is in Appendix A.3.

4 Implementation Algorithm

4.1 Wasserstein Instantiation

To obtain a tractable learning objective, we instantiate the discrepancy measures $\mathcal{D}_\omega$, $\mathcal{D}_X$, and $\mathcal{D}_S$ with the Wasserstein distance (Wiesemann et al., 2014; Esfahani and Kuhn, 2015; Sinha et al., 2018).

Definition 1 (Wasserstein Distance) *Let P and Q be probability measures on a measurable space $\mathcal{Z}$ equipped with a cost function $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$. The Wasserstein distance induced by c is*

$$\mathcal{W}_c(P, Q) := \inf_{\xi \in \Gamma(P,Q)} \int_{\mathcal{Z}\times\mathcal{Z}} c(u, u')\, \mathrm{d}\xi(u, u'), \tag{10}$$

where $\Gamma(P, Q)$ denotes the set of all couplings of P and Q.

At the global level, we model uncertainty in the group mixture weights over the finite set $\mathcal{G} = \{1, \ldots, m\}$. Since the groups have no natural geometry, we equip $\mathcal{G}$ with the discrete cost $d_G(g, g') = \mathbf{1}[g \ne g']$. Under this cost, the Wasserstein distance reduces to the total variation distance, i.e., $\mathcal{W}_{d_G}(\boldsymbol{\omega}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\omega} - \boldsymbol{\alpha}\|_1$. At the local level, we model within-group uncertainty through Wasserstein balls:

$$\mathcal{W}_{c_X}(Q_g^X, P_g^X) \le \rho_g^X, \qquad \mathcal{W}_{c_S}\big(Q_g^S(\cdot \mid \boldsymbol{x}, \boldsymbol{v}), P_g^S(\cdot \mid \boldsymbol{x}, \boldsymbol{v})\big) \le \rho_g^S.$$

Here, c_X is the transport cost on the prompt space and c_S is the label-flip cost on $\{+1, -1\}$. The radii ρ_g^X and ρ_g^S independently control prompt-level and preference-level robustness within group g.

4.2 Tractable Objective

In this section, we derive a tractable reformulation by successively robustifying the conditional preference distribution, the prompt marginal distribution, and the distribution over group weights.

Preference-level robustification. Recall that, for group g, the conditional distribution factorizes as $P_g(V, S \mid X) = P_g^V(V \mid X)\, P_g^S(S \mid X, V)$. Since the candidate response pair V is observed and fixed in the offline dataset, the remaining preference-level uncertainty lies in the conditional distribution over binary labels $P_g^S(S \mid X, V)$. Thus, we model preference-level shift as a perturbation of this binary label distribution, holding the response pair fixed. In standard DPO data, each comparison is encoded with $S = +1$, so $P_g^S(\cdot \mid \boldsymbol{x}, \boldsymbol{v}) = \delta_{+1}$. For each group g and each prompt and response pair $(\boldsymbol{x}, \boldsymbol{v}) = (\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)$, we define

$$\mathcal{U}_{g,\boldsymbol{x},\boldsymbol{v}}^S = \Big\{ q \in \Delta(\{+1, -1\}) : \mathcal{W}_{c_S}(q, \delta_{+1}) \le \rho_g^S \Big\}, \tag{11}$$

with the label-flip cost $c_S(s, s') = \mathbf{1}[s \neq s']$. Under this cost, $\mathcal{W}_{c_S}(q, \delta_{+1})$ equals the probability of $s = -1$ under q, so the ball in Eq. (11) allows at most a fraction ρ_g^S of the label mass to be flipped from $+1$ to -1. The worst-case preference-level robust loss is then

$$\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) := \sup_{q \in \mathcal{U}_{g,\boldsymbol{x},\boldsymbol{v}}^S} \mathbb{E}_{S \sim q}\big[\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2, S; \boldsymbol{\theta})\big]. \tag{12}$$

Proposition 1 (Closed-form preference-level robust loss) *For any group g, prompt $\boldsymbol{x}$ and candidate response pair $(\boldsymbol{y}_1, \boldsymbol{y}_2)$, we have*

$$\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) = \varphi\big(m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})\big) + \rho_g^S \big[m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})\big]_+, \tag{13}$$

where $[t]_+ = \max\{t, 0\}$. The proof is provided in Appendix A.1.

Proposition 2 (Optimal robust preference margin) *For any group g, let $M = m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})$ denote the DPO margin defined in Eq.* (1)*. Then $\mathcal{L}_g(M) = \varphi(M) + \rho_g^S [M]_+$ is minimized at*

$$M_g^\star = \begin{cases} +\infty, & \rho_g^S = 0, \\ \log \dfrac{1 - \rho_g^S}{\rho_g^S}, & 0 < \rho_g^S < \frac{1}{2}, \\ 0, & \rho_g^S \ge \frac{1}{2}. \end{cases} \tag{14}$$

The proof is provided in Appendix A.2.

Remark 1 *Proposition 1 shows that preference-level Wasserstein robustness yields a robustified loss that appends a one-sided margin penalty scaled by ρ_g^S. Proposition 2 indicates that, unlike standard DPO, preference-level Wasserstein robustness does not drive the margin to grow unboundedly. Instead, it yields a finite optimal target margin. Moreover, this target margin decreases as ρ_g^S increases, implying that stronger uncertainty in the conditional preference distribution yields a more conservative optimum and reduces overconfidence in weak or potentially misaligned preferences.*

Prompt-level robustification. We next robustify the prompt marginal distribution. For each group g, the local robust risk is

$$\mathcal{R}_g(\boldsymbol{\theta}) := \sup_{Q_g^X \in \mathcal{U}_g^X} \mathbb{E}_{X \sim Q_g^X,\, V \sim P_g^V(\cdot \mid X)}\big[\psi_g(X, V; \boldsymbol{\theta})\big], \tag{15}$$

where $\psi_g(\boldsymbol{x}, \boldsymbol{v}; \boldsymbol{\theta})$ is the worst-case preference loss in Eq. (8). Under the label-flip Wasserstein ball in Eq. (11), ψ_g admits the closed form $\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})$ given in Proposition 1.

To obtain a tractable empirical formulation, we follow the robust NLP fine-tuning literature and perturb prompts in the embedding space rather than the discrete token space (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2020). Let $\boldsymbol{h} = \mathcal{E}(\boldsymbol{x}) \in \mathbb{R}^d$ denote the prompt embedding and $\boldsymbol{z}_i = (\boldsymbol{h}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-)$ the i-th embedded observation. For group g, the empirical distribution over embedded preference observations is $\widehat{P}_g = \frac{1}{n_g} \sum_{i \in I_g} \delta_{\boldsymbol{z}_i}$.

Following Wasserstein DRO with local input perturbations (Kwon et al., 2020) and embedding-space adversarial training in NLP (Zhu et al., 2020; Miyato et al., 2017; Jiang et al., 2020), we perturb the prompt embedding of each preference observation while keeping the associated response pair fixed. Accordingly, we define the transport cost

$$c_X\big((\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-), (\boldsymbol{h}', \tilde{\boldsymbol{y}}^+, \tilde{\boldsymbol{y}}^-)\big) = \begin{cases} \frac{1}{2}\|\boldsymbol{h} - \boldsymbol{h}'\|_2^2, & (\boldsymbol{y}^+, \boldsymbol{y}^-) = (\tilde{\boldsymbol{y}}^+, \tilde{\boldsymbol{y}}^-), \\ +\infty, & \text{otherwise.} \end{cases} \tag{16}$$

For this empirical approximation, we define the local robust risk as

$$\widehat{\mathcal{R}}_g(\boldsymbol{\theta}) = \sup_{\widetilde{Q}_g^X : \mathcal{W}_{c_X}(\widetilde{Q}_g^X, \widehat{P}_g) \le \rho_g^X} \mathbb{E}_{(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-) \sim \widetilde{Q}_g^X} \big[\bar{\mathcal{L}}_g(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})\big]. \tag{17}$$

Lemma 1 (Wasserstein DRO duality (Sinha et al., 2018; Zhang et al., 2025)) *Under standard regularity conditions for Wasserstein DRO duality, for any reference distribution Q, radius $\rho \ge 0$, loss ℓ, and transport cost c, we have*

$$\sup_{P:\, \mathcal{W}_c(P,Q) \le \rho} \mathbb{E}_P[\ell(Z)] = \inf_{\gamma \ge 0} \Big\{ \gamma\rho + \mathbb{E}_Q\big[\ell_p(Z;\gamma)\big] \Big\}, \tag{18}$$

where $\ell_p(u_0;\gamma) = \sup_u \{\ell(u) - \gamma\, c(u, u_0)\}$ is the Wasserstein dual surrogate loss.

Applying Lemma 1 to Eq. (17) with $\ell = \bar{\mathcal{L}}_g$ and $c = c_X$ yields the dual form

$$\widehat{\mathcal{R}}_g(\boldsymbol{\theta}) = \inf_{\gamma_g \ge 0} \left\{ \gamma_g \rho_g^X + \frac{1}{n_g} \sum_{i \in I_g} \ell_{\boldsymbol{\theta}}^g(\boldsymbol{z}_i; \gamma_g) \right\}, \tag{19}$$

where $\boldsymbol{z}_i = (\boldsymbol{h}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-)$. Under the transport cost in Eq. (16), changing the response pair incurs infinite cost, so the inner maximization only searches over perturbations of the prompt embedding:

$$\ell_{\boldsymbol{\theta}}^g(\boldsymbol{z}_i; \gamma_g) = \sup_{\boldsymbol{\delta} \in \mathbb{R}^d} \left\{ \bar{\mathcal{L}}_g(\boldsymbol{h}_i + \boldsymbol{\delta}, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2}\|\boldsymbol{\delta}\|_2^2 \right\}. \tag{20}$$

Final empirical objective. Combining the global Wasserstein reweighting with the dual form of the local risks, the final empirical TPO objective is

$$\min_{\boldsymbol{\theta} \in \Theta} \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum\nolimits_{g=1}^m \omega_g\, \widehat{\mathcal{R}}_g(\boldsymbol{\theta}), \tag{21}$$

where $\widehat{\mathcal{R}}_g(\boldsymbol{\theta})$ is given in Eq. (19) and $\bar{\mathcal{L}}_g$ in Eq. (13). Eq. (21) reduces the abstract hierarchical Wasserstein robust objective to a concrete min-max training problem.

4.3 Optimization Algorithm

While the dual reformulation in Eq. (20) is exact, explicitly optimizing over γ_g at each training step is computationally prohibitive. Following standard practice in Wasserstein robust optimization (Sinha et al., 2018; Zhang et al., 2019; Kwon et al., 2020; Bai et al., 2025), we treat γ_g as a fixed hyperparameter and additionally constrain each perturbation to $\mathcal{B}_g^X = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2^2 \le 2\rho_g^X\}$. For each group g, the resulting penalty-based surrogate objective is

$$\widetilde{\mathcal{R}}_g(\boldsymbol{\theta}; \gamma_g) := \gamma_g \rho_g^X + \frac{1}{n_g} \sum\nolimits_{i \in I_g} \sup_{\boldsymbol{\delta} \in \mathcal{B}_g^X} \left[\bar{\mathcal{L}}_g(\boldsymbol{h}_i + \boldsymbol{\delta}, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2}\|\boldsymbol{\delta}\|_2^2 \right]. \tag{22}$$

The full training procedure is summarized in Algorithm 1. Following alternating optimization strategies in adversarial training (Madry et al., 2018; Sinha et al., 2018; Jiao et al., 2025a) and robust group reweighting (Sagawa et al., 2020), each iteration proceeds in three stages. **(i) Adversarial Perturbation.** For each embedded preference observation $(\boldsymbol{h}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-)$ in the mini-batch, we approximate the prompt-level supremum in Eq. (22) by running K steps of projected gradient ascent on $\boldsymbol{\delta}$ with the fixed quadratic penalty. **(ii) Group-weight update.** Using the perturbed embeddings from stage (i), we compute group-wise stochastic estimates $\widetilde{\mathcal{R}}_g^{(t)}$ and perform one step of exponentiated gradient ascent projected onto $\mathcal{U}^\omega$, which implements the worst-case group reweighting over the L_1-constrained simplex. **(iii) Policy update.** We update the policy parameters $\boldsymbol{\theta}$ by taking a gradient descent step on the weighted surrogate objective $\sum_{g=1}^m \omega_g^{(t)} \widetilde{\mathcal{R}}_g^{(t)}$, with the adversarial perturbations and group weights fixed.

Algorithm 1 Alternating Optimization for TPO

Input: Step sizes $\eta_\theta, \eta_\omega, \eta_\delta$; penalty parameters γ_g, inner steps K, and total iterations T.
for $t = 1, \ldots, T$ **do**
  Sample group mini-batches $\mathcal{B}^{(t)} = \{\mathcal{B}_g^{(t)}\}_{g=1}^m$, $\mathcal{B}_g^{(t)} \subseteq \mathcal{D}_g$
  for each $g \in \mathcal{G}$ and $(\boldsymbol{h}_i, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-) \in \mathcal{B}_g^{(t)}$ **do**
    Initialize $\boldsymbol{\delta}_{g,i}^{(0)} \leftarrow \mathbf{0}$
    for $k = 0, \ldots, K-1$ **do**

$$\boldsymbol{\delta}_{g,i}^{(k+1)} \leftarrow \Pi_{\mathcal{B}_g^X}\left(\boldsymbol{\delta}_{g,i}^{(k)} + \eta_\delta \nabla_{\boldsymbol{\delta}}\left[\bar{\mathcal{L}}_g\left(\boldsymbol{h}_i + \boldsymbol{\delta}, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-; \boldsymbol{\theta}^{(t-1)}\right) - \tfrac{\gamma_g}{2}\|\boldsymbol{\delta}\|_2^2\right]_{\boldsymbol{\delta}=\boldsymbol{\delta}_{g,i}^{(k)}}\right)$$

    end for
  end for
  for each $g \in \mathcal{G}$ **do**

$$\widetilde{\mathcal{R}}_g^{(t)}(\boldsymbol{\theta}) \leftarrow \gamma_g \rho_g^X + \frac{1}{|\mathcal{B}_g^{(t)}|}\sum\nolimits_{i \in \mathcal{B}_g^{(t)}}\left[\bar{\mathcal{L}}_g\left(\boldsymbol{h}_i + \boldsymbol{\delta}_{g,i}^{(K)}, \boldsymbol{y}_i^+, \boldsymbol{y}_i^-; \boldsymbol{\theta}\right) - \tfrac{\gamma_g}{2}\|\boldsymbol{\delta}_{g,i}^{(K)}\|_2^2\right]$$

  end for

$$\boldsymbol{\omega}^{(t)} \leftarrow \Pi_{\mathcal{U}^\omega}\left(\boldsymbol{\omega}^{(t-1)} \odot \exp\left(\eta_\omega \widetilde{\mathcal{R}}^{(t)}(\boldsymbol{\theta}^{(t-1)})\right)\right)$$

$$\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)} - \eta_\theta \ \nabla_{\boldsymbol{\theta}}\left(\sum\nolimits_{g=1}^m \omega_g^{(t)} \widetilde{\mathcal{R}}_g^{(t)}(\boldsymbol{\theta})\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1)}}$$

end for
Return: $\boldsymbol{\theta}^{(T)}$

Table 2: Average and worst-group preference accuracy (%) under different OOD shifts. The best result is in **bold** and the second-best is underlined.

Qwen-2.5-0.5B	ID		Group		Prompt		Preference		Combined		Avg. Rank
	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	
DPO	86.87	77.78	81.80	76.15	83.17	76.98	82.93	78.16	82.60	76.40	5.30
Dr.DPO	84.81	81.36	81.65	77.06	82.75	76.84	82.24	75.86	81.27	75.22	6.50
WDPO	84.87	84.13	82.72	78.90	85.08	82.20	84.01	81.46	84.21	78.76	3.20
TPO (Group)	84.90	77.27	82.96	76.76	83.46	78.28	82.31	79.24	82.06	76.11	5.25
TPO (Prompt)	85.93	83.77	82.84	80.43	84.69	81.05	85.02	**83.03**	84.37	79.65	2.40
TPO (Preference)	85.64	78.28	82.77	77.98	84.30	79.75	83.28	79.99	83.24	76.11	4.25
TPO	**86.95**	**85.12**	**84.61**	**81.65**	**85.47**	**82.56**	**85.44**	82.34	**85.25**	**79.94**	**1.10**
LLaMA-3.2-1B	ID		Group		Prompt		Preference		Combined		Avg. Rank
	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	
DPO	88.75	86.19	88.23	83.44	87.81	85.43	86.91	**81.42**	87.44	82.61	4.15
Dr.DPO	84.16	81.00	84.65	79.87	83.73	80.98	82.87	78.87	85.51	79.13	7.00
WDPO	87.53	84.17	87.24	80.86	87.78	85.82	85.28	79.73	86.71	81.16	5.80
TPO (Group)	89.03	87.02	88.61	83.39	88.32	85.86	86.94	81.22	87.54	82.90	3.50
TPO (Prompt)	89.78	86.37	88.50	81.19	88.71	85.41	87.30	81.31	87.63	82.03	3.75
TPO (Preference)	89.75	**87.55**	89.00	83.17	89.00	86.23	87.40	81.24	87.87	82.61	2.60
TPO	**90.57**	**87.55**	**89.11**	**83.50**	**89.14**	**86.70**	**88.25**	81.31	**88.31**	**83.19**	**1.20**

5 Experiments

5.1 Experimental Setup

We evaluate TPO on three tasks: OOD preference alignment, preference poisoning, and downstream generalization. For OOD preference alignment and downstream generalization, we compare against standard DPO and several recent robust baselines: GRPO (Ramesh et al., 2024), Dr.DPO (Wu et al., 2025), and WDPO (Xu et al., 2025). We also include three variants of TPO: TPO (Group), TPO (Prompt), and TPO (Preference), which activate only the group-level, prompt-level, and preference-level robustness components, respectively. Note that TPO (Group) reduces to GRPO (Ramesh et al., 2024) as a special case. For preference poisoning, we compare against rDPO (Chowdhury et al., 2024) and Dr.DPO (Wu et al., 2025), two robust DPO baselines designed to mitigate noisy preferences.

OOD Preference Alignment. Following prior emotion-based evaluations of preference shift (Xu et al., 2025; Sahu and Wells, 2025), we construct a benchmark from the Emotion dataset (Saravia et al., 2018) with four shift environments: **(1) Group-level shift:** we partition samples into six latent groups via K-means clustering on prompt and response embeddings, and change the group mixture weights at test time; **(2) Prompt-level shift:** we apply synonym substitution to prompts to simulate prompt shift; **(3) Preference-level shift:** we construct two reward functions r_1 and r_2 and simulate preference

Table 3: Performance comparison on Open LLM Leaderboard v2, which evaluates LLMs across six tasks: MMLU-Pro, BBH, GPQA, MATH-Hard, IFEval, and MuSR.

Task	LLaMA-3.2-1B				LLaMA-3.2-3B				LLaMA-3.1-8B			
	DPO	TPO			DPO	TPO			DPO	TPO		
		$\rho_g^S = 0$	$\rho_g^S = 0.01$	$\rho_g^S = 0.05$		$\rho_g^S = 0$	$\rho_g^S = 0.01$	$\rho_g^S = 0.05$		$\rho_g^S = 0$	$\rho_g^S = 0.005$	$\rho_g^S = 0.01$
MMLU-Pro	20.0	20.5	20.5	20.2	31.2	31.5	31.5	31.7	36.8	36.7	37.1	37.1
BBH	35.6	35.8	35.9	35.6	45.2	45.6	45.9	45.7	51.1	50.9	51.7	51.3
GPQA	23.2	23.5	23.7	23.3	28.0	28.4	28.3	28.1	30.0	30.2	30.8	30.1
MATH-Hard	8.1	9.4	8.3	8.2	21.4	21.1	20.1	21.0	19.5	20.9	20.7	19.3
IFEval	51.3	50.8	53.0	53.1	61.5	63.5	63.6	65.6	75.4	77.6	79.6	77.5
MuSR	31.6	31.6	31.2	31.0	37.5	37.7	37.4	37.1	38.0	38.1	38.0	37.4

shift via the convex interpolation $r_\alpha = \alpha r_1 + (1-\alpha) r_2$, which shifts the conditional preference distribution $P_g^S(S \mid V, X)$; and **(4) Combined shift:** all three shifts are applied simultaneously. We use Qwen2.5-0.5B-Instruct (Qwen Team, 2024) and LLaMA-3.2-1B-Instruct (Grattafiori et al., 2024) as base models, and report average and worst-group preference accuracy (Wu et al., 2025) to measure overall performance and worst-case robustness, respectively. Further details on reward model training, data construction, OOD shift settings, and hyperparameter selection are provided in Appendix B.1.

Preference Poisoning Attack Robustness. To evaluate robustness against training-time preference poisoning, we follow established poisoning protocols in RLHF (Wang et al., 2024b; Pathmanathan et al., 2025; Kusaka et al., 2026) and construct poisoned training sets by flipping the preference labels (i.e., swapping the preferred and rejected responses) of a fraction of the training pairs, while keeping the test set clean. We consider two attack strategies: **(1) Random preference flip attack** (Pathmanathan et al., 2025), where poisoned pairs are selected uniformly at random; and **(2) Margin-based preference flip attack** (Pathmanathan et al., 2025), where the adversary targets the pairs with the largest margins computed by the reward model used for preference construction. We use LLaMA-3.2-1B-Instruct (Grattafiori et al., 2024) as the base model and evaluate all methods on clean test preferences. Further details on poisoned data construction, hyperparameter selection, and evaluation settings are provided in Appendix B.2.

Downstream Generalization. To assess the downstream generalization of TPO under potential distribution shift, we evaluate on the Open LLM Leaderboard v2 (Hugging Face, 2024). We fine-tune LLaMA-3.2-1B-Instruct, LLaMA-3.2-3B-Instruct, and LLaMA-3.1-8B-Instruct for one epoch. Following the pipeline of Meng et al. (2024), we collect on-policy data annotated by the ArmoRM reward model (Wang et al., 2024a). Full training and evaluation details are provided in Appendix B.3.

5.2 Experimental Results

OOD Preference Alignment Results. As shown in Table 2, TPO achieves the best average accuracy across all evaluation environments on both backbones, while attaining competitive worst-group accuracy in most settings. While the single-level variants TPO (Group), TPO (Prompt), and TPO (Preference) each provide partial gains, their combination in TPO yields stronger overall performance across both metrics and nearly all shift environments, suggesting the complementary benefits of the three robustness levels.

Preference Poisoning Attack Robustness Results. As shown in Figure 1, all methods degrade as the poisoning rate increases. Under random preference flip attacks, rDPO performs best as expected, given its explicit debiasing design for random label flips. TPO is not specifically designed for this setting, yet generally outperforms DPO and remains competitive with Dr.DPO. Under margin-based flip attacks, Dr.DPO performs strongly at low poisoning rates, but TPO surpasses all baselines at higher poisoning rates. Unlike rDPO and Dr.DPO, TPO explicitly penalizes large positive preference margins, which may mitigate overconfidence under preference uncertainty and contribute to its stronger performance against margin-based poisoning at higher poisoning rates.

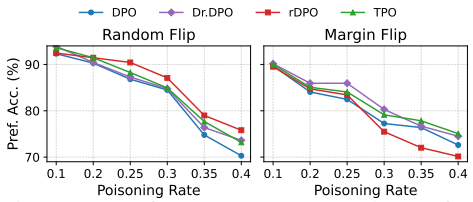


Figure 1: Preference accuracy under preference poisoning.

Downstream Generalization Results. As shown in Table 3, TPO achieves comparable or better performance over DPO on most benchmarks across the three model scales, suggesting that the proposed robustness mechanisms largely preserve general capabilities. Gains are most pronounced on IFEval and also appear on BBH, while performance on GPQA and MMLU-Pro remains largely stable. We also observe that different preference-level robustness radii lead to varying trade-offs across tasks, which is expected given the heterogeneous nature of the leaderboard benchmarks. Interestingly, the improvements tend to become larger on IFEval as model scale increases, suggesting that larger models may benefit more from the proposed robustness mechanisms on instruction-following tasks. Overall, these results suggest that TPO can improve downstream generalization while largely retaining broad task performance.

6 Related Work

Robust Preference Optimization. DPO simplifies preference alignment by directly optimizing over pairwise data (Rafailov et al., 2023; Azar et al., 2024), but remains vulnerable to distribution shift at deployment (Ramesh et al., 2024; Wu et al., 2025). Recent robust DPO methods address this from two angles: WDPO and KLDPO (Xu et al., 2025) robustify DPO using ambiguity sets over the overall preference distribution, while rDPO (Chowdhury et al., 2024) and Dr.DPO (Wu et al., 2025) target preference noise. Another line of work focuses on instability in preference optimization, including likelihood displacement (Luo et al., 2026; Razin et al., 2025; Chen et al., 2024), likelihood underdetermination (Guo et al., 2025), and catastrophic preference shift (Jian et al., 2026). Group robust methods such as GRPO (Ramesh et al., 2024) and pluralistic alignment work (Sorensen et al., 2024; Yao et al., 2025; Tao et al., 2024) motivate modeling preference heterogeneity across diverse user populations. However, none of these methods provide a unified robust objective that simultaneously captures group-level shift, prompt-level shift, and conditional preference shift.

Distributionally Robust Optimization. Classical robust optimization models uncertainty through structured uncertainty sets (Yang et al., 2008), while distributionally robust optimization (DRO) considers ambiguity sets over probability distributions to address distributional uncertainty (Jiao et al., 2022, 2026; Esfahani and Kuhn, 2015; Wiesemann et al., 2014). In particular, Wasserstein DRO (Sinha et al., 2018; Esfahani and Kuhn, 2015) admits tractable dual reformulations and connects naturally to adversarial perturbations in embedding spaces (Madry et al., 2018; Zhu et al., 2020; Jiang et al., 2020). For heterogeneous data, group DRO (Sagawa et al., 2020) optimizes worst-case group performance but treats each group distribution as fixed, leaving within-group shifts unaddressed (Koh et al., 2021; Chowdhury et al., 2024). Recent work on hierarchical ambiguity sets (Jo et al., 2026) argues that group-level robustness alone is insufficient. Our work brings this perspective to preference optimization via a hierarchically structured rectangular ambiguity set that further decomposes intra-group uncertainty into prompt-level and preference-level shifts, thereby providing a more complete and interpretable robust alignment framework.

7 Conclusion

We proposed tri-level distributionally robust preference optimization (TPO), which addresses OOD preference alignment by jointly modeling three levels of distribution shift through a hierarchically structured rectangular ambiguity set. By instantiating it with Wasserstein distances, we obtain an objective that combines worst-case group reweighting, prompt-level adversarial perturbation, and a closed-form preference-level margin penalty. Theoretical analysis establishes finite-sample convergence guarantees, and comprehensive experiments demonstrate the effectiveness of our method.

Acknowledgments and Disclosure of Funding

The work of Chengtao Jian and Kai Yang was supported in part by NSFC under Grant 12371519 and in part by Shenzhen Loop Area Institute under Project FP202602. The authors declare no competing interests.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Rémi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 238, pages 4447–4455, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Xingjian Bai, Guangyi He, Yifan Jiang, and Jan Obłój. Wasserstein distributional adversarial training for deep neural networks. *CoRR*, abs/2502.09352, 2025. URL https://arxiv.org/abs/2502.09352.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. URL https://jmlr.org/papers/v3/bartlett02a.html.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit S. Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=8tzjEMF0Vq.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: aligning language models with noisy feedback. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=yhpDKSw7yA.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2023. URL https://arxiv.org/abs/2310.01377.

Steven Diamond and Stephen Boyd. CVXPY: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115 – 166, 2015.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

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

Kaiyang Guo, Yinchuan Li, and Zhitang Chen. Proximalized preference optimization for diverse feedback types: A decomposed perspective on dpo. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hugging Face. Open LLM leaderboard 2, 2024. URL https://huggingface.co/collections/open-llm-leaderboard/open-llm-leaderboard-2.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.

Chengtao Jian, Kai Yang, and Yang Jiao. Tri-Level Navigator: LLM-Empowered Tri-Level Learning for Time Series OOD Generalization. In *Advances in Neural Information Processing Systems*, volume 37, pages 110613–110642, 2024. doi: 10.52202/079017-3511.

Chengtao Jian, Kai Yang, Tianhao Gao, Wuguang Ni, Keying Yang, Bowen Xiao, Jiajun Liu, and Ye Ouyang. Stable preference optimization: A bilevel approach to catastrophic preference shift. *arXiv preprint arXiv:2507.07723*, 2026.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, 2020. URL https://aclanthology.org/2020.acl-main.197/.

Yang Jiao, Kai Yang, and Dongjin Song. Distributed distributionally robust optimization with non-convex objectives. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=AUJT3rj2F5U.

Yang Jiao, Kai Yang, Tiancheng Wu, Dongjin Song, and Chengtao Jian. Asynchronous distributed bilevel optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

Yang Jiao, Kai Yang, Tiancheng Wu, Chengtao Jian, and Jianwei Huang. Provably convergent federated trilevel learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12928–12937, 2024.

Yang Jiao, Xiaodong Wang, and Kai Yang. PR-Attack: Coordinated prompt-RAG attacks on retrieval-augmented generation in large language models via bilevel optimization. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 656–667. Association for Computing Machinery, 2025a. doi: 10.1145/3726302.3730058.

Yang Jiao, Kai Yang, and Chengtao Jian. Dtzo: Distributed trilevel zeroth order learning with provable non-asymptotic convergence. In *International Conference on Machine Learning*, pages 27873–27916, 2025b.

Yang Jiao, Kai Yang, and Dongjin Song. Federated distributionally robust optimization with non-convex objectives: Algorithm and analysis. *IEEE Transactions on Mobile Computing*, 25(1): 1219–1235, 2026.

Sung Ho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets, 2026. URL https://openreview.net/forum?id=mruL1LDjzV.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran S. Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, volume 139, pages 5637–5664, 2021. URL http://proceedings.mlr.press/v139/koh21a.html.

Shigeki Kusaka, Keita Saito, Mikoto Kudo, Takumi Tanabe, Akifumi Wachi, and Youhei Akimoto. Cost-minimized label-flipping poisoning attack to LLM alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37538–37546, 2026. doi: 10.1609/aaai.v40i44.41087. URL https://ojs.aaai.org/index.php/AAAI/article/view/41087.

Yongchan Kwon, Wonyoung Kim, Joong-Ho Won, and Myunghee Cho Paik. Principled learning method for wasserstein distributionally robust optimization with local perturbations. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 5567–5576, 2020. URL http://proceedings.mlr.press/v119/kwon20a.html.

Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. Robust prompt optimization for large language models against distribution shifts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1539–1554, 2023. URL https://aclanthology.org/2023.emnlp-main.95/.

Yan Li and Alexander Shapiro. Rectangularity and duality of distributionally robust markov decision processes. *Mathematical Programming*, 2025. doi: 10.1007/s10107-025-02297-y. URL https://doi.org/10.1007/s10107-025-02297-y.

Ya Liu, Kai Yang, Yu Zhu, Keying Yang, Chengtao Jian, Wuguang Ni, Xiaozhou Ye, and Ye Ouyang. A nested zeroth-order fine-tuning approach for cloud-edge llm agents. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 3–15. Springer, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Haocheng Luo, Zehang Deng, Thanh-Toan Do, Mehrtash Harandi, Dinh Phung, and Trung Le. Sharpness-aware minimization in logit space efficiently enhances direct preference optimization. In *International Conference on Learning Representations*, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1X3g2_xl.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, and Furong 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, pages 27556–27564, 2025. doi: 10.1609/aaai.v39i26.34968. URL https://ojs.aaai.org/index.php/AAAI/article/view/34968.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. In *Advances in Neural Information Processing Systems*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5e1c255653eb98cef13f45b2d337c882-Abstract-Conference.html.

Rui Qiao, Zhaoxuan Wu, Jingtan Wang, Pang Wei Koh, and Bryan Kian Hsiang Low. Group-robust sample reweighting for subpopulation shifts via influence functions. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aQj9Ifxrl6.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou-Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. In *Advances in Neural Information Processing Systems*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/4147dfaa46cd7e20a2aecb91097ae8cc-Abstract-Conference.html.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sharan Sahu and Martin T Wells. Online distributionally robust llm alignment via regression to relative reward. *arXiv preprint arXiv:2509.19104*, 2025.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, 2018. URL https://aclanthology.org/D18-1404/.

Parker Seegmiller and Sarah Masud Preum. Measuring distribution shift in user prompts and its effects on llm performance, 2026. URL https://arxiv.org/abs/2604.17650.

Alexander Shapiro. Rectangular sets of probability measures. *Operations Research*, 64(2):528–541, 2016. doi: 10.1287/opre.2015.1466.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=gQpBnRHwxM.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346, 2024. doi: 10.1093/pnasnexus/pgae346.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. ISBN 9781108231596. doi: 10.1017/9781108231596. URL https://doi.org/10.1017/9781108231596.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. ISBN 9781108627771. doi: 10.1017/9781108627771. URL https://doi.org/10.1017/9781108627771.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 10582–10592, 2024a. URL https://aclanthology.org/2024.findings-emnlp.620/.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. RLHFPoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2551–2570, 2024b. URL https://aclanthology.org/2024.acl-long.140/.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Oper. Res.*, 62:1358–1376, 2014.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CbfsKHiWEn.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=D19hc2XPeZ.

Kai Yang, Yihong Wu, Jianwei Huang, Xiaodong Wang, and Sergio Verdú. Distributed robust optimization for communication networks. In *IEEE INFOCOM 2008*, pages 1157–1165. IEEE, 2008. doi: 10.1109/INFOCOM.2008.171. URL https://doi.org/10.1109/INFOCOM.2008.171.

Kai Yang, Jianwei Huang, Yihong Wu, Xiaodong Wang, and Mung Chiang. Distributed robust optimization (DRO), Part I: Framework and example. *Optimization and Engineering*, 15(1): 35–67, March 2014. doi: 10.1007/s11081-012-9198-y. URL https://doi.org/10.1007/s11081-012-9198-y.

Keying Yang, Hao Wang, Chengtao Jian, and Kai Yang. Stcor: A trilevel syllogism-driven reasoning framework. In *Proceedings of the 13th Workshop on Argument Mining and Reasoning*, pages 1–4, 2026.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No preference left behind: Group distributional preference optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bgpNJBD6Va.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the International Conference on Machine Learning*, volume 97, pages 7472–7482, 2019. URL http://proceedings.mlr.press/v97/zhang19p.html.

Luhao Zhang, Jincheng Yang, and Rui Gao. A short and general duality proof for wasserstein distributionally robust optimization. *Operations Research*, 73(4):2146–2155, 2025. ISSN 0030-364X. doi: 10.1287/opre.2023.0135. URL https://doi.org/10.1287/opre.2023.0135.

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

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygzbyHFvB.

A Proofs

A.1 Proof of Proposition 1

Recall that, for a fixed group g and a fixed prompt and response pair $(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)$, the preference-level robust loss is defined as

$$\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) = \sup_{q \in \mathcal{U}_g^S} \mathbb{E}_{S \sim q}\big[\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2, S; \boldsymbol{\theta})\big], \tag{23}$$

where

$$\mathcal{U}_g^S = \left\{q \in \Delta(\{+1, -1\}) : \mathcal{W}_{c_S}(q, \delta_{+1}) \le \rho_g^S\right\}, \qquad c_S(s, s') = \mathbf{1}[s \neq s']. \tag{24}$$

Here, δ_{+1} corresponds to the canonicalized observed preference label $S = +1$, obtained by ordering the candidate pair as $(\boldsymbol{y}_1, \boldsymbol{y}_2) = (\boldsymbol{y}^+, \boldsymbol{y}^-)$.

Since the label space contains only two points, any $q \in \Delta(\{+1, -1\})$ can be parameterized as $q(+1) = 1 - \varepsilon$, where $q(-1) = \varepsilon$ for some $\varepsilon \in [0, 1]$. Under the discrete label-flip cost c_S, the Wasserstein distance from q to δ_{+1} is exactly the mass assigned to the flipped label $\mathcal{W}_{c_S}(q, \delta_{+1}) = \varepsilon$. Therefore, assuming $0 \le \rho_g^S \le 1$, the Wasserstein constraint is equivalent to $0 \le \varepsilon \le \rho_g^S$.

Let $M = m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})$. Using the labeled DPO loss $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2, S; \boldsymbol{\theta}) = \varphi(S \cdot M)$, we have

$$\begin{aligned}\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) &= \sup_{0 \le \varepsilon \le \rho_g^S} \Big[(1 - \varepsilon)\varphi(M) + \varepsilon\, \varphi(-M)\Big] \\ &= \varphi(M) + \sup_{0 \le \varepsilon \le \rho_g^S} \varepsilon\big(\varphi(-M) - \varphi(M)\big).\end{aligned} \tag{25}$$

For $\varphi(t) = \log(1 + e^{-t})$, we have $\varphi(-t) - \varphi(t) = t$. Thus, we can rewrite the preference-level robust loss as

$$\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) = \varphi(M) + \sup_{0 \le \varepsilon \le \rho_g^S} \varepsilon M. \tag{26}$$

Since the last term is linear in ε, its supremum is attained at $\varepsilon = \rho_g^S$ when $M \ge 0$, and at $\varepsilon = 0$ when $M < 0$. Therefore, we have

$$\sup_{0 \le \varepsilon \le \rho_g^S} \varepsilon M = \rho_g^S [M]_+. \tag{27}$$

Substituting this back, we obtain

$$\bar{\mathcal{L}}_g(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta}) = \varphi\big(m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})\big) + \rho_g^S \big[m(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2; \boldsymbol{\theta})\big]_+, \tag{28}$$

which completes the proof of Proposition 1.

A.2 Proof of Proposition 2

Let $\mathcal{L}_g(M) = \varphi(M) + \rho_g^S [M]_+$ and $\varphi(M) = \log(1 + e^{-M})$. We analyze the minimizer over $M \in \mathbb{R}$. First, when $\rho_g^S = 0$, the objective reduces to $\mathcal{L}_g(M) = \varphi(M)$. Since $\varphi'(M) = -\frac{1}{1+e^M} < 0$, $\varphi(M)$ is strictly decreasing on $\mathbb{R}$. Therefore, its infimum is approached as $M \to +\infty$, and the optimal target margin is $M_g^\star = +\infty$.

We now consider the case $\rho_g^S > 0$. Since $[M]_+$ is piecewise linear, we analyze the cases $M < 0$ and $M > 0$ separately. For $M < 0$, we have $[M]_+ = 0$, hence, we obtain $\mathcal{L}_g(M) = \varphi(M)$. Its derivative satisfies $\mathcal{L}_g'(M) = \varphi'(M) = -\frac{1}{1+e^M} < 0$. Therefore, $\mathcal{L}_g(M)$ is strictly decreasing on $(-\infty, 0)$.

For $M > 0$, we have $[M]_+ = M$, and thus

$$\mathcal{L}_g(M) = \varphi(M) + \rho_g^S M. \tag{29}$$

Its derivative is

$$\mathcal{L}_g'(M) = -\frac{1}{1 + e^M} + \rho_g^S. \tag{30}$$

Setting $\mathcal{L}'_g(M) = 0$ yields

$$\frac{1}{1+e^M} = \rho_g^S \quad \Longleftrightarrow \quad e^M = \frac{1-\rho_g^S}{\rho_g^S}, \tag{31}$$

so any stationary point on $(0, \infty)$ must satisfy

$$M = \log \frac{1-\rho_g^S}{\rho_g^S}. \tag{32}$$

This stationary point is positive if and only if $0 < \rho_g^S < \frac{1}{2}$. When $0 < \rho_g^S < \frac{1}{2}$, the derivative $\mathcal{L}'_g(M)$ changes sign from negative to positive at

$$M_g^\star = \log \frac{1-\rho_g^S}{\rho_g^S}, \tag{33}$$

so this point is the global minimizer. When $\rho_g^S \geq \frac{1}{2}$, the stationary point above is non-positive. Moreover, for all $M > 0$, we have

$$\mathcal{L}'_g(M) = -\frac{1}{1+e^M} + \rho_g^S > 0, \tag{34}$$

so $\mathcal{L}_g(M)$ is increasing on $(0, \infty)$. Since it is decreasing on $(-\infty, 0)$, the global minimizer is attained at the boundary $M = 0$.

Therefore, we have shown that the optimal target margin is

$$M_g^\star \in \arg\min_{M \in \mathbb{R}} \mathcal{L}_g(M) = \begin{cases} +\infty, & \rho_g^S = 0, \\ \log \dfrac{1-\rho_g^S}{\rho_g^S}, & 0 < \rho_g^S < \frac{1}{2}, \\ 0, & \rho_g^S \geq \frac{1}{2}, \end{cases} \tag{35}$$

which proves Proposition 2.

A.3 Proof of Theorem 1

Theorem 1 analyzes the population and empirical minimizers of the fixed-penalty Wasserstein dual objective, with each penalty parameter $\gamma_g > 0$ held fixed. Throughout this subsection, $\mathcal{R}_g$ and $\mathcal{R}_{\text{TPO}}$ refer to this fixed-penalty objective, without minimization over γ_g. For a sample, the DPO margin is defined as

$$m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) = \beta \left(\log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+ \mid \boldsymbol{h})}{\pi_{\text{ref}}(\boldsymbol{y}^+ \mid \boldsymbol{h})} - \log \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^- \mid \boldsymbol{h})}{\pi_{\text{ref}}(\boldsymbol{y}^- \mid \boldsymbol{h})} \right). \tag{36}$$

The logistic loss is denoted by $\varphi(t) = \log(1 + e^{-t})$. The preference-level robust loss for group g is

$$\bar{\mathcal{L}}_g(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) = \varphi(m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})) + \rho_g^S [m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})]_+. \tag{37}$$

The fixed-penalty local robust sample loss is

$$\ell_{\boldsymbol{\theta}}^g(U; \gamma_g) = \sup_{\boldsymbol{\delta} \in \mathbb{R}^d} \left[\bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2} \|\boldsymbol{\delta}\|_2^2 \right]. \tag{38}$$

The population and empirical local robust risk for group g are

$$\mathcal{R}_g(\boldsymbol{\theta}) = \gamma_g \rho_g^X + \mathbb{E}_{P_g}\left[\ell_{\boldsymbol{\theta}}^g(U; \gamma_g)\right], \quad \widehat{\mathcal{R}}_g(\boldsymbol{\theta}) = \gamma_g \rho_g^X + \frac{1}{n_g} \sum_{i=1}^{n_g} \ell_{\boldsymbol{\theta}}^g(U_{g,i}; \gamma_g). \tag{39}$$

The hierarchical population and empirical TPO objectives are defined as

$$\mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum_{g=1}^m \omega_g \mathcal{R}_g(\boldsymbol{\theta}), \quad \widehat{\mathcal{R}}_{\text{TPO}}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum_{g=1}^m \omega_g \widehat{\mathcal{R}}_g(\boldsymbol{\theta}). \tag{40}$$

Throughout the proof, we assume that samples are independent within each group, $U_{g,1}, \ldots, U_{g,n_g} \overset{\text{i.i.d.}}{\sim} P_g$.

Lemma 2 (Boundedness and Lipschitzness of the DPO margin) *Suppose Assumption 1 holds. Then, for each original or adversarially perturbed prompt embedding $\tilde{\boldsymbol{h}}$, each response pair $(\boldsymbol{y}^+, \boldsymbol{y}^-)$, and each $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$,*

$$|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})| \le M_\beta, \qquad M_\beta := 4\beta B C_\phi, \tag{41}$$

and

$$|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}')| \le 2\beta C_\phi \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2. \tag{42}$$

Proof. Under the log-linear policy class, we have

$$\log \pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+ \mid \tilde{\boldsymbol{h}}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{y}^- \mid \tilde{\boldsymbol{h}}) = \boldsymbol{\theta}^\top \big(\phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+) - \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^-)\big). \tag{43}$$

Since π_{ref} belongs to the same class with parameter $\boldsymbol{\theta}_{\mathrm{ref}}$, we have

$$m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) = \beta \left\langle \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+) - \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^-), \boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ref}} \right\rangle. \tag{44}$$

By Assumption 1, we have

$$\|\phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+) - \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^-)\|_2 \le 2C_\phi, \qquad \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ref}}\|_2 \le 2B. \tag{45}$$

Therefore,

$$|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})| \le \beta \cdot 2C_\phi \cdot 2B = 4\beta B C_\phi = M_\beta. \tag{46}$$

So we can obtain

$$\begin{aligned} &|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}')| \\ &= \beta \left| \left\langle \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+) - \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^-), \boldsymbol{\theta} - \boldsymbol{\theta}' \right\rangle \right| \\ &\le \beta \|\phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+) - \phi(\tilde{\boldsymbol{h}}, \boldsymbol{y}^-)\|_2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \\ &\le 2\beta C_\phi \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{aligned} \tag{47}$$

which completes the proof.

Lemma 3 (Boundedness and Lipschitzness of $\bar{\mathcal{L}}_g$) *Suppose Assumption 1 holds. Then, for each group g, each original or adversarially perturbed prompt embedding $\tilde{\boldsymbol{h}}$, each response pair $(\boldsymbol{y}^+, \boldsymbol{y}^-)$, and each $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have*

$$0 \le \bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) \le K_g,$$

where $K_g = \log(1+e^{M_\beta}) + \rho_g^S M_\beta$ and $M_\beta = 4\beta B C_\phi$. Moreover, $\bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})$ is L_g-Lipschitz, and $L_g = 2\beta C_\phi(1 + \rho_g^S)$.

Proof. By definition,

$$\bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \theta) = \varphi(m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \theta)) + \rho_g^S [m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \theta)]_+. \tag{48}$$

Since $\varphi(t) \ge 0$, $[t]_+ \ge 0$, and $\rho_g^S \ge 0$, we have $\bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) \ge 0$.

According to Lemma 2, we have $|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})| \le M_\beta$. Hence, we obtain

$$\varphi\Big(m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})\Big) \le \log(1 + e^{M_\beta}), \tag{49}$$

$$\Big[m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})\Big]_+ \le M_\beta. \tag{50}$$

Combining (49) and (50), we get

$$\bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) \le \log(1 + e^{M_\beta}) + \rho_g^S M_\beta = K_g. \tag{51}$$

For Lipschitzness, $\varphi(t) = \log(1 + e^{-t})$ is 1-Lipschitz because $|\varphi'(t)| = \frac{1}{1+e^t} \le 1$. And $t \mapsto [t]_+$ is also 1-Lipschitz. Therefore, $f(t) = \varphi(t) + \rho_g^S [t]_+$ is $(1 + \rho_g^S)$-Lipschitz. Combining this with Lemma 2, we have

$$\begin{aligned} &\Big|\bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \bar{\mathcal{L}}_g(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}')\Big| \\ &\le (1 + \rho_g^S) \Big|m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - m(\tilde{\boldsymbol{h}}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}')\Big| \\ &\le 2\beta C_\phi (1 + \rho_g^S) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \end{aligned} \tag{52}$$

which completes the proof.

Lemma 4 (Rademacher complexity of the local robust loss class) *Let $\mathcal{F}_g = \{\ell^g_{\boldsymbol{\theta}}(\cdot;\gamma_g) : \boldsymbol{\theta} \in \Theta\}$, where $\gamma_g > 0$ is fixed. Under Assumption 1, the class $\mathcal{F}_g$ satisfies*

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \le A_g \sqrt{\frac{d}{n_g}},$$

where $A_g = CK_g\sqrt{\log\left(e + \frac{3BL_g}{K_g}\right)}$, $C > 0$ is a universal numerical constant, $K_g = \log(1 + e^{M_\beta}) + \rho_g^S M_\beta$, $M_\beta = 4\beta B C_\phi$, and $L_g = 2\beta C_\phi(1 + \rho_g^S)$.

Proof. We first prove that the fixed-penalty local robust loss $\ell^g_{\boldsymbol{\theta}}(U;\gamma_g)$ is bounded and Lipschitz in $\boldsymbol{\theta}$. By Lemma 3, for every admissible perturbation $\boldsymbol{\delta}$, we have $0 \le \bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) \le K_g$.

Since $-\frac{\gamma_g}{2}\|\boldsymbol{\delta}\|^2 \le 0$, we can obtain

$$\ell^g_{\boldsymbol{\theta}}(U;\gamma_g) = \sup_{\boldsymbol{\delta}} \left[\bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2}\|\boldsymbol{\delta}\|_2^2\right] \le K_g. \tag{53}$$

Moreover, because $\boldsymbol{\delta} = 0$ is admissible, $\ell^g_{\boldsymbol{\theta}}(U;\gamma_g) \ge \bar{\mathcal{L}}_g(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) \ge 0$. Thus, we have $0 \le \ell^g_{\boldsymbol{\theta}}(U;\gamma_g) \le K_g$.

For Lipschitzness, we define

$$\mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\delta}) = \bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2}\|\boldsymbol{\delta}\|_2^2. \tag{54}$$

Then, we have

$$\begin{aligned}
|\ell^g_{\boldsymbol{\theta}}(U;\gamma_g) - \ell^g_{\boldsymbol{\theta}'}(U;\gamma_g)| &= \left|\sup_{\boldsymbol{\delta}} \mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\delta}) - \sup_{\boldsymbol{\delta}} \mathcal{J}_{\boldsymbol{\theta}'}(\boldsymbol{\delta})\right| \\
&\le \sup_{\boldsymbol{\delta}} |\mathcal{J}_{\boldsymbol{\theta}}(\boldsymbol{\delta}) - \mathcal{J}_{\boldsymbol{\theta}'}(\boldsymbol{\delta})| \\
&= \sup_{\boldsymbol{\delta}} \left|\bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \bar{\mathcal{L}}_g(\boldsymbol{h} + \boldsymbol{\delta}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}')\right| \\
&\le L_g\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.
\end{aligned} \tag{55}$$

Hence, $\ell^g_{\boldsymbol{\theta}}$ is L_g-Lipschitz in $\boldsymbol{\theta}$ with $L_g = 2\beta C_\phi(1 + \rho_g^S)$.

Thus, an $\frac{\varepsilon}{L_g}$-cover of Θ under $\|\cdot\|_2$ induces an ε-cover of $\mathcal{F}_g$ under $\|\cdot\|_\infty$. Indeed, for any $\boldsymbol{\theta} \in \Theta$, choose $\boldsymbol{\theta}_j$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \le \frac{\varepsilon}{L_g}$. Then

$$\|\ell^g_{\boldsymbol{\theta}}(\cdot;\gamma_g) - \ell^g_{\boldsymbol{\theta}_j}(\cdot;\gamma_g)\|_\infty \le L_g\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \le \varepsilon. \tag{56}$$

Hence, we can get

$$N(\varepsilon, \mathcal{F}_g, \|\cdot\|_\infty) \le N(\frac{\varepsilon}{L_g}, \Theta, \|\cdot\|_2). \tag{57}$$

By Dudley's entropy integral for Rademacher complexity (Wainwright, 2019), the entropy bound above gives

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \le \frac{C}{\sqrt{n_g}} \int_0^{K_g} \sqrt{\log N(\varepsilon, \mathcal{F}_g, \|\cdot\|_\infty)}\, d\varepsilon. \tag{58}$$

By the standard volumetric covering bound (Vershynin, 2018; Wainwright, 2019), we have that

$$\log N(\varepsilon, \mathcal{F}_g, \|\cdot\|_\infty) \le d\log\left(1 + \frac{3BL_g}{\varepsilon}\right) \le d\log\left(e + \frac{3BL_g}{\varepsilon}\right) \tag{59}$$

So we obtain

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \le \frac{C}{\sqrt{n_g}} \int_0^{K_g} \sqrt{d\log\left(e + \frac{3BL_g}{\varepsilon}\right)}\, d\varepsilon. \tag{60}$$

Let $R_g = 3BL_g$. It remains to bound the following integral:

$$I_g = \int_0^{K_g} \sqrt{\log\left(e + \frac{R_g}{\varepsilon}\right)}\, d\varepsilon. \tag{61}$$

By setting $u = \frac{\varepsilon}{K_g}$ and $a_g = \frac{R_g}{K_g}$, we can rewrite this integral as

$$I_g = K_g \int_0^1 \sqrt{\log\left(e + \frac{a_g}{u}\right)}\, du. \tag{62}$$

For $u \in (0, 1]$, we have $e + \frac{a_g}{u} \leq \frac{e+a_g}{u}$, and hence

$$\log\left(e + \frac{a_g}{u}\right) \leq \log(e + a_g) + \log(1/u). \tag{63}$$

Using $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, we obtain

$$I_g \leq K_g \sqrt{\log(e + a_g)} + K_g \int_0^1 \sqrt{\log(1/u)}\, du. \tag{64}$$

Moreover,

$$\int_0^1 \sqrt{\log(1/u)}\, du = \int_0^\infty s^{1/2} e^{-s}\, ds = \Gamma(3/2) = \frac{\sqrt{\pi}}{2}. \tag{65}$$

Since $\log(e + a_g) \geq 1$, the fixed numerical term $\frac{\sqrt{\pi}}{2}$ can be absorbed into a universal constant. Therefore,

$$I_g \leq CK_g \sqrt{\log(e + a_g)} = CK_g \sqrt{\log\left(e + \frac{3BL_g}{K_g}\right)}. \tag{66}$$

Substituting this into Eq. (60), we obtain

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \leq CK_g \sqrt{\log\left(e + \frac{3BL_g}{K_g}\right)} \sqrt{\frac{d}{n_g}}. \tag{67}$$

By the definition of $A_g = CK_g \sqrt{\log\left(e + \frac{3BL_g}{K_g}\right)}$, we obtain

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \leq A_g \sqrt{\frac{d}{n_g}}. \tag{68}$$

This completes the proof of Lemma 4.

Lemma 5 (Group-wise local uniform convergence) *Suppose Assumption 1 holds. Then, with probability at least $1 - \delta$, for all $g \in [m]$, we have*

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta}) \right| \leq \epsilon_g,$$

where $\epsilon_g = 2A_g \sqrt{\frac{d}{n_g}} + K_g \sqrt{\frac{\log(2m/\delta)}{2n_g}}$.

Proof. For a fixed group g, the constant term $\gamma_g \rho_g^X$ appears in both $\mathcal{R}_g(\boldsymbol{\theta})$ and $\widehat{\mathcal{R}}_g(\boldsymbol{\theta})$, so it cancels and we have

$$\mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta}) = \mathbb{E}_{P_g}[\ell^g_{\boldsymbol{\theta}}(U; \gamma_g)] - \frac{1}{n_g} \sum_{i=1}^{n_g} \ell^g_{\boldsymbol{\theta}}(U_{g,i}; \gamma_g). \tag{69}$$

By the standard Rademacher uniform convergence bound (Bartlett and Mendelson, 2002) for functions bounded in $[0, K_g]$, with probability at least $1 - \delta_g$,

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta}) \right| \leq 2\mathfrak{R}_{n_g}(\mathcal{F}_g) + K_g \sqrt{\frac{\log(2/\delta_g)}{2n_g}}. \tag{70}$$

Using Lemma 4,

$$\mathfrak{R}_{n_g}(\mathcal{F}_g) \leq A_g \sqrt{\frac{d}{n_g}}. \tag{71}$$

Setting $\delta_g = \frac{\delta}{m}$ and applying a union bound over $g = 1, \ldots, m$ gives the result.

Lemma 6 (From local convergence to hierarchical convergence) *Suppose the event in Lemma 5 holds and $\mathcal{U}^\omega \subseteq \Delta^{m-1}$. Then*

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta})\right| \le \epsilon_{\mathrm{H}},$$

where $\epsilon_{\mathrm{H}} = \sup_{\boldsymbol{\omega}\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \epsilon_g$. In particular, $\epsilon_{\mathrm{H}} \le \max_{g\in[m]} \epsilon_g$.

Proof. Fix $\boldsymbol{\theta} \in \Theta$. Using

$$\left|\sup_x \mathcal{A}_{\boldsymbol{\theta}}(x) - \sup_x \mathcal{A}_{\boldsymbol{\theta}'}(x)\right| \le \sup_x |\mathcal{A}_{\boldsymbol{\theta}}(x) - \mathcal{A}_{\boldsymbol{\theta}'}(x)|, \tag{72}$$

we obtain

$$\begin{aligned}
&\left|\mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta})\right| \\
&= \left|\sup_{\omega\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \mathcal{R}_g(\boldsymbol{\theta}) - \sup_{\omega\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \widehat{\mathcal{R}}_g(\boldsymbol{\theta})\right| \\
&\le \sup_{\omega\in\mathcal{U}^\omega} \left|\sum_{g=1}^m \omega_g \left(\mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta})\right)\right| \\
&\le \sup_{\omega\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \left|\mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta})\right|.
\end{aligned} \tag{73}$$

By Lemma 5, for all $g \in [m]$ and $\boldsymbol{\theta} \in \Theta$, we have

$$\left|\mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta})\right| \le \epsilon_g \tag{74}$$

Therefore, we can get

$$\left|\mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta})\right| \le \sup_{\omega\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \epsilon_g = \epsilon_{\mathrm{H}}. \tag{75}$$

Since $\mathcal{U}^\omega \subseteq \Delta^{m-1}$, we have $\omega_g \ge 0$ and $\sum_g \omega_g = 1$. Hence,

$$\sup_{\omega\in\mathcal{U}^\omega} \sum_{g=1}^m \omega_g \epsilon_g \le \max_{g\in[m]} \epsilon_g. \tag{76}$$

Lemma 7 (Strong convexity of the population TPO objective) *Suppose Assumptions 1 and 2 hold. Let $M_\beta = 4\beta B C_\phi$ and $\kappa_\beta = \frac{\beta^2 e^{M_\beta}}{(1+e^{M_\beta})^2}$. Then the population risk $\mathcal{R}_{\mathrm{TPO}}$ is $\kappa_\beta\lambda$-strongly convex over Θ. Consequently, for $\boldsymbol{\theta}^\star \in \arg\min_{\boldsymbol{\theta}\in\Theta} \mathcal{R}_{TPO}(\boldsymbol{\theta})$, we have*

$$\mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}) - \mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star) \ge \frac{\kappa_\beta\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2$$

Proof. For a fixed sample $U = (\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-)$, let $z(U) = \phi(\boldsymbol{h}, \boldsymbol{y}^+) - \phi(\boldsymbol{h}, \boldsymbol{y}^-)$. Under the log-linear policy class, we have

$$m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) = \beta\langle z(U), \boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ref}}\rangle. \tag{77}$$

The Hessian of the DPO logistic loss $\varphi(m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ is

$$\nabla^2_{\boldsymbol{\theta}} \varphi(m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})) = \beta^2 \sigma(m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}))\sigma(-m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}))z(U)z(U)^\top. \tag{78}$$

By Lemma 2, we have

$$|m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})| \le M_\beta. \tag{79}$$

Since

$$\sigma(t)\sigma(-t) = \frac{e^t}{(1+e^t)^2}, \tag{80}$$

for all $|t| \le M_\beta$, we can obtain

$$\sigma(t)\sigma(-t) \ge \frac{e^{M_\beta}}{(1+e^{M_\beta})^2}. \tag{81}$$

Therefore, we have

$$\nabla^2_{\boldsymbol{\theta}} \varphi(m(\boldsymbol{h}, \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta})) \succeq \kappa_\beta z(U) z(U)^\top. \tag{82}$$

We now establish strong convexity of the hierarchical fixed-penalty objective. For $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ and $t \in [0,1]$, let $\boldsymbol{\theta}_t = t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2$. Choose worst-case group weights $\boldsymbol{\omega}_t \in \mathcal{U}^\omega$ and measurable optimal perturbations $T_{g,\boldsymbol{\theta}_t}(U)$ at $\boldsymbol{\theta}_t$.

We define

$$F_t(\boldsymbol{\theta}) = \sum_{g=1}^m \omega_{t,g} \left[\gamma_g \rho_g^X + \mathbb{E}_{P_g}\left[\bar{\mathcal{L}}_g(\boldsymbol{h} + T_{g,\boldsymbol{\theta}_t}(U), \boldsymbol{y}^+, \boldsymbol{y}^-; \boldsymbol{\theta}) - \frac{\gamma_g}{2} \|T_{g,\boldsymbol{\theta}_t}(U)\|_2^2 \right] \right]. \tag{83}$$

By the preceding Hessian bound and Assumption 2, the logistic-loss component of F_t has curvature at least $\kappa_\beta \lambda$. The preference-level penalty $\rho_g^S [m]_+$ is convex in $\boldsymbol{\theta}$, and the perturbation penalty is independent of $\boldsymbol{\theta}$. Thus, F_t is $\kappa_\beta \lambda$-strongly convex.

By optimality of the selected weights and perturbations, we have $F_t(\boldsymbol{\theta}_t) = \mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}_t)$, and

$$F_t(\boldsymbol{\theta}) \le \mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta.$$

Consequently,

$$\begin{aligned} \mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}_t) \le t\mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}_1) + (1-t)\mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}_2) \\ - \frac{\kappa_\beta \lambda}{2} t(1-t)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2. \end{aligned} \tag{84}$$

Hence, $\mathcal{R}_{\text{TPO}}$ is $\kappa_\beta \lambda$-strongly convex. Since $\boldsymbol{\theta}^\star$ minimizes this objective over the convex set Θ, we have

$$\mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}) - \mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}^\star) \ge \frac{\kappa_\beta \lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2^2. \tag{85}$$

Theorem 1 (Finite-sample convergence of TPO) *Let $M_\beta = 4\beta B C_\phi$ and $\kappa_\beta = \frac{\beta^2 e^{M_\beta}}{(1+e^{M_\beta})^2}$. Suppose Assumptions 1 and 2 hold. With probability at least $1-\delta$, we have*

$$\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \le 2\sqrt{\frac{\max_{g\in[m]}\left[2A_g\sqrt{\frac{d}{n_g}} + K_g\sqrt{\frac{\log(2m/\delta)}{2n_g}}\right]}{\kappa_\beta \lambda}},$$

where $K_g = \log(1+e^{M_\beta}) + \rho_g^S M_\beta$, $L_g = 2\beta C_\phi(1+\rho_g^S)$, $A_g = C K_g \sqrt{\log\big(e + 3BL_g/K_g\big)}$, and $C > 0$ is a universal constant. The proof is in Appendix A.3.

Proof. Now we are ready to prove Theorem 1. By Lemma 5, with probability at least $1-\delta$, simultaneously for all $g \in [m]$,

$$\sup_{\theta \in \Theta} \left| \mathcal{R}_g(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_g(\boldsymbol{\theta}) \right| \le \epsilon_g, \tag{86}$$

where

$$\epsilon_g = 2A_g\sqrt{\frac{d}{n_g}} + K_g\sqrt{\frac{\log(2m/\delta)}{2n_g}}. \tag{87}$$

By Lemma 6, we have

$$\sup_{\boldsymbol{\theta} \in \Theta} \left| \mathcal{R}_{\text{TPO}}(\boldsymbol{\theta}) - \widehat{\mathcal{R}}_{\text{TPO}}(\boldsymbol{\theta}) \right| \le \epsilon_{\text{H}}, \tag{88}$$

where

$$\epsilon_{\text{H}} = \sup_{\boldsymbol{\omega} \in \mathcal{U}^\omega} \sum_{g=1}^m \omega_g \epsilon_g. \tag{89}$$

Since $\widehat{\boldsymbol{\theta}}$ minimizes the empirical TPO objective, we have

$$\widehat{\mathcal{R}}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}}) \le \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star). \tag{90}$$

Therefore,

$$\begin{aligned}\mathcal{R}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star) &= \Big[\mathcal{R}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}}) - \widehat{\mathcal{R}}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}})\Big] \\ &\quad + \Big[\widehat{\mathcal{R}}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}}) - \widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star)\Big] \\ &\quad + \Big[\widehat{\mathcal{R}}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star) - \mathcal{R}_{\mathrm{TPO}}(\boldsymbol{\theta}^\star)\Big] \\ &\le \epsilon_{\mathrm{H}} + 0 + \epsilon_{\mathrm{H}} = 2\epsilon_{\mathrm{H}}.\end{aligned} \tag{91}$$

By Lemma 7,

$$\mathcal{R}_{\mathrm{TPO}}(\widehat{\boldsymbol{\theta}}) - \mathcal{R}_{\mathrm{TPO}}(\theta^\star) \ge \frac{\kappa_\beta \lambda}{2}\|\widehat{\theta} - \theta^\star\|_2^2. \tag{92}$$

Combining the two inequalities gives

$$\frac{\kappa_\beta \lambda}{2}\|\widehat{\theta} - \theta^\star\|_2^2 \le 2\epsilon_{\mathrm{H}}. \tag{93}$$

Thus,

$$\|\widehat{\theta} - \theta^\star\|_2 \le 2\sqrt{\frac{\epsilon_{\mathrm{H}}}{\kappa_\beta \lambda}}. \tag{94}$$

Finally, since $\mathcal{U}^\omega \subseteq \Delta^{m-1}$, we have

$$\epsilon_{\mathrm{H}} = \sup_{\omega \in \mathcal{U}^\omega} \sum_{g=1}^{m} \omega_g \epsilon_g \le \max_{g \in [m]} \epsilon_g. \tag{95}$$

Substituting the expression for ϵ_g yields the stated max-form bound.

B Experimental Details

B.1 Details for OOD preference alignment

We construct a controlled benchmark from the `dair-ai/emotion` dataset (Saravia et al., 2018), which consists of short English social-media texts labeled with six emotion categories. We remove the underrepresented `surprise` class, retaining five emotions: `sadness`, `joy`, `love`, `anger`, and `fear`. The benchmark is built in three steps: we first train a multi-objective emotion reward model, then use it to construct preference pairs, and finally define latent groups and the OOD evaluation environments.

Multi-objective reward model. Following Xu et al. (2025), we convert the dataset into a multi-label reward dataset by randomly concatenating one to three source texts and merging their emotion labels into a five-dimensional binary vector $z \in \{0,1\}^5$. We fine-tune a Qwen2.5-0.5B-Instruct backbone with a linear classification head, trained with binary cross-entropy loss

$$\mathcal{L}_{\mathrm{RM}} = -\sum\nolimits_{k=1}^{5} \big[z_k \log \sigma(h_k(x)) + (1 - z_k)\log(1 - \sigma(h_k(x)))\big],$$

where $h(x) \in \mathbb{R}^5$ denotes the logits produced by the classification head applied to the hidden state of the last non-padding token. Training uses learning rate 5×10^{-5}, weight decay 0.01, dropout 0.1, and maximum sequence length 128 for 8 epochs.

Base model and preference pair construction. We fine-tune two base models: Qwen2.5-0.5B-Instruct and LLaMA-3.2-1B-Instruct via supervised fine-tuning (SFT) on the original emotion texts for 10 epochs with learning rate 5×10^{-6}, weight decay 0.01, max gradient norm 10. The resulting models serve as both the initial policy and the reference model for preference optimization. For each prompt (first four tokens of each source text), we sample eight candidate completions with temperature 0.7, top-$p = 0.9$, max 64 tokens, and score them with a scalar preference reward defined as a convex combination of two emotion rewards: $R_\alpha(y) = \alpha\, r_{\texttt{anger}}(y) + (1-\alpha)\, r_{\texttt{fear}}(y)$. Training pairs are constructed under $\alpha_{\mathrm{train}} = 0.2$ by selecting the highest- and lowest-scoring completions as chosen and rejected responses, respectively, and filtering out pairs with reward margin below 0.02.

Group Construction and OOD Shift Settings. Since the Emotion dataset lacks explicit domain annotations, we construct six latent groups by applying K-means clustering ($K = 6$) to the prompt and response embeddings extracted from Qwen2.5-0.5B-Instruct. We consider four OOD evaluation settings corresponding to the distribution shifts described in Section 3.1: **(i) Group-level shift:** We change the group mixture proportions from a skewed training distribution $(40\%, 25\%, 15\%, 10\%, 7\%, 3\%)$ to a uniform test distribution. **(ii) Prompt-level shift:** We apply synonym substitution to at most one word per prompt from a predefined vocabulary (`feel`, `happy`, `sad`, `angry`, `afraid`, `love`). **(iii) Preference-level shift:** We change the reward mixing coefficient from $\alpha_{\text{train}} = 0.2$ to $\alpha_{\text{test}} = 0.8$, shifting the preference criterion from `fear` toward `anger`. **(iv) Combined shift:** We apply all three shifts simultaneously. Table 4 summarizes the evaluation settings.

Table 4: Environment construction for the Emotion preference benchmark.

Environment	Group Proportions	Prompt Perturbation	Reward Mixture α
`train`	Skewed	None	0.2
`val`	Skewed	None	0.2
`id_test`	Skewed	None	0.2
`group_test`	Uniform	None	0.2
`prompt_test`	Skewed	Synonym substitution	0.2
`preference_test`	Skewed	None	0.8
`combined_test`	Uniform	Synonym substitution	0.8

Hyperparameter Search. We set the prompt adversarial perturbation step size $\eta_\delta = 0.001$, the policy learning rate $\eta_\theta = 1 \times 10^{-6}$, global batch size 64, inner adversarial steps $K = 1$, and scaling coefficient $\beta = 0.1$. All models are optimized with AdamW (Loshchilov and Hutter, 2019) and trained on 2 NVIDIA H100 80GB GPUs. We select the remaining hyperparameters of TPO via grid search on the validation set over the ranges in Table 5, using average preference accuracy on the validation set as the selection criterion. For baselines, we use the same validation metric and tune over comparable candidate sets following their original papers when applicable. Otherwise, we adopt the reported defaults.

Table 5: Hyperparameter search ranges.

Hyperparameter	Search Range
γ_g	$\{0.1,\ 1.0\}$
η_ω	$\{0.01,\ 0.03,\ 0.05,\ 0.1\}$
ρ^ω	$\{0.2, 0.1, 0.01,\ 0.05\}$
ρ_g^X	$\{0.01,\ 0.001,\ 0.0005,\ 0.0001\}$
ρ_g^S	$\{0.2, 0.1,\ 0.01,\ 0.005,\ 0.001\}$

B.2 Details for Preference Poisoning Attack

Benchmark construction. For a selected poisoned example $(\boldsymbol{x}, \boldsymbol{y}^+, \boldsymbol{y}^-)$, poisoning replaces it with $(\boldsymbol{x}, \boldsymbol{y}^-, \boldsymbol{y}^+)$. We construct poisoned training sets from the controlled Emotion preference benchmark (Appendix B.1), swapping chosen and rejected responses for a selected subset of training pairs at poisoning rates $\{0.1, 0.2, 0.25, 0.3, 0.35, 0.4\}$, while validation and test splits remain clean. We consider two attack strategies described in Section 5.

Training and evaluation setup. All methods use learning rate $\eta_\theta = 5 \times 10^{-6}$, batch size 64, $\beta = 0.1$, and AdamW (Loshchilov and Hutter, 2019) on 2 NVIDIA H100 80GB GPUs. To ensure a comparable tuning budget, we select method-specific robustness hyperparameters on the clean validation split. For rDPO, we tune $\epsilon \in \{0.01, 0.1, 0.2\}$. For Dr.DPO, we tune $\tau \in \{0.1, 0.5, 1.0\}$. For TPO, we disable prompt-level and group-level robustness and tune the preference-level radius $\rho_g^S \in \{0.05, 0.1, 0.2\}$. All selected models are trained on poisoned preference data and evaluated on

Table 6: Average and worst-group accuracy (%) on the Emotion preference benchmark.

Qwen-2.5-0.5B	ID		Group		Prompt		Preference		Combined	
	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc
DPO	86.87±0.32	77.78±0.87	81.80±0.15	76.15±0.92	83.17±0.46	76.98±1.10	82.93±0.49	78.16±1.99	82.60±0.92	76.40±0.51
Dr.DPO	84.81±0.23	81.36±1.02	81.65±0.35	77.06±0.78	82.75±0.69	76.84±1.22	82.24±0.44	75.86±1.65	81.27±0.94	75.22±0.73
WDPO	84.87±0.29	84.13±1.18	82.72±0.16	78.90±0.49	85.08±0.22	82.20±2.36	84.01±0.65	81.46±1.83	84.21±0.31	78.76±1.14
TPO (Group)	84.90±0.32	77.27±1.52	82.96±0.23	76.76±1.40	83.46±0.32	78.28±0.87	82.31±0.22	79.24±0.12	82.06±0.23	76.11±0.88
TPO (Prompt)	85.93±0.55	83.77±0.76	82.84±0.26	80.43±2.12	84.69±0.70	81.05±2.11	85.02±0.97	83.03±1.61	84.37±0.39	79.65±0.88
TPO (Preference)	85.64±0.22	78.28±1.75	82.77±0.09	77.98±0.92	84.30±0.43	79.75±0.71	83.28±0.10	79.99±0.74	83.24±0.34	76.11±0.88
TPO	86.95±0.70	85.12±0.24	84.61±0.99	81.65±0.62	85.47±0.86	82.56±1.32	85.44±1.13	82.34±2.84	85.25±0.78	79.94±1.84

LLaMA-3.2-1B	ID		Group		Prompt		Preference		Combined	
	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc
DPO	88.75±0.06	86.19±0.15	88.23±0.38	83.44±1.92	87.81±0.27	85.43±0.83	86.91±0.22	81.42±2.43	87.44±0.71	82.61±0.87
Dr.DPO	84.16±0.72	81.00±1.24	84.65±0.57	79.87±2.29	83.73±0.62	80.98±1.31	82.87±1.26	78.87±2.21	85.51±0.66	79.13±1.74
WDPO	87.53±1.03	84.17±1.77	87.24±1.05	80.86±0.57	87.78±1.03	85.82±1.71	85.28±0.96	79.73±1.82	86.71±1.31	81.16±1.00
TPO (Group)	89.03±0.39	87.02±0.68	88.61±0.33	83.39±2.12	88.32±0.33	85.86±1.10	86.94±0.21	81.22±2.31	87.54±0.38	82.90±0.50
TPO (Prompt)	89.78±1.32	86.37±1.03	88.50±1.01	81.19±0.99	88.71±0.93	85.41±1.00	87.30±0.52	81.31±1.75	87.63±0.17	82.03±1.00
TPO (Preference)	89.75±0.33	87.55±0.43	89.00±0.74	83.17±1.71	89.00±0.16	86.23±0.06	87.40±0.54	81.24±2.43	87.87±0.30	82.61±0.00
TPO	90.57±0.27	87.55±0.43	89.11±0.50	83.50±1.14	89.14±0.47	86.70±0.74	88.25±0.34	81.31±0.87	88.31±0.30	83.19±0.50

the clean test split. For both the random and margin-based preference flip attacks, we report average preference accuracy to evaluate robustness to preference poisoning.

B.3 Details for Downstream Generalization

Dataset and preference construction. We use UltraFeedback (Cui et al., 2023) as the preference training source. It contains instruction-following prompts together with multiple model-generated responses and fine-grained quality annotations. Following Meng et al. (2024), we first collect on-policy responses from the SFT model, and then construct candidate response pairs using the ArmoRM reward model (Wang et al., 2024a) by selecting the highest- and lowest-scoring responses for each prompt. We use the naturally defined UltraFeedback categories as the group structure for robust preference alignment.

Training and evaluation setup. We fine-tune all models for one epoch using the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate $\eta_\theta = 5 \times 10^{-6}$, global batch size 64, and $\beta = 0.1$, on 2 NVIDIA H100 80GB GPUs. For TPO, the preference-level radius is selected from $\rho_g^S \in \{0.005, 0.01, 0.05, 0.1, 0.2\}$. We evaluate on Open LLM Leaderboard v2 (Hugging Face, 2024), which comprises six benchmark families: IFEval, BBH, MATH-Hard, GPQA, MuSR, and MMLU-Pro.

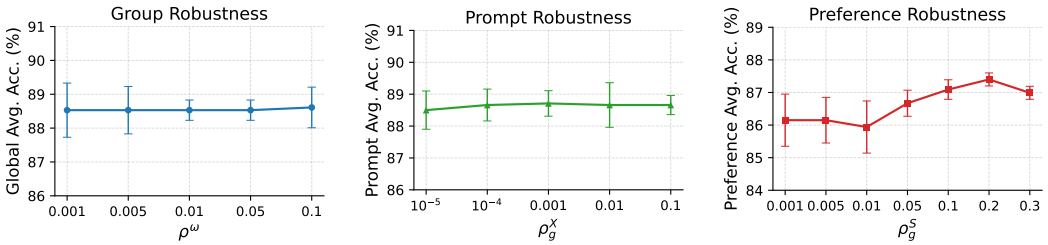
B.4 Additional Implementation Details

Following embedding-space adversarial training (Zhu et al., 2020), we apply the prompt perturbation $\boldsymbol{\delta}$ to the prompt-token embedding matrix, leaving response tokens unchanged. The same perturbation is shared across the chosen, rejected, and reference branches, with padding positions masked out. The reference policy is evaluated on the perturbed prompts with its parameters frozen. We use the quadratic penalty $\frac{\gamma_g}{2}\|\boldsymbol{\delta}\|_F^2$ and project perturbations onto $\|\boldsymbol{\delta}\|_F^2 \le 2\rho_g^X$, where $\|\cdot\|_F$ denotes the Frobenius norm.

For group-weight updates, we adopt exponentiated gradient ascent (Sagawa et al., 2020; Ramesh et al., 2024), followed by a KL projection onto the group-mixture ambiguity set. Given group-wise surrogate losses $\widetilde{\mathcal{R}}_g^{(t)}$, we compute provisional weights $\widetilde{\omega}_g \propto \omega_g^{(t-1)} \exp(\eta_\omega \widetilde{\mathcal{R}}_g^{(t)})$ and update

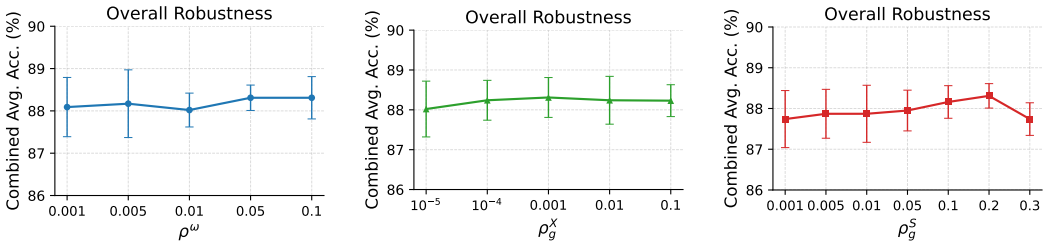
$$\boldsymbol{\omega}^{(t)} = \arg \min_{\substack{\boldsymbol{\omega} \in \Delta^{m-1} \\ \|\boldsymbol{\omega} - \boldsymbol{\alpha}\|_1 \le 2\rho^\omega}} \mathcal{D}_{\mathrm{KL}}(\boldsymbol{\omega} \,\|\, \widetilde{\boldsymbol{\omega}}). \tag{96}$$

The ℓ_1 constraint follows from $\mathcal{W}_1^G(\boldsymbol{\omega}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\omega} - \boldsymbol{\alpha}\|_1$ under the discrete group transport cost. We solve this convex projection using `CVXPY` (Diamond and Boyd, 2016).



(a) Group-level robustness. (b) Prompt-level robustness. (c) Preference-level robustness.

Figure 2: Sensitivity of group-, prompt-, and preference-level robustness to their corresponding radii.



(a) Group radius. (b) Prompt radius. (c) Preference radius.

Figure 3: Sensitivity of overall robustness to the three robustness radii.

C Additional Experiments

Detailed results of Emotion benchmark. The main paper reports mean results. Here we additionally report the standard deviations across three runs with different random seeds. The results are shown in Table 6.

Hyperparameter sensitivity analysis. We further study the sensitivity of TPO to the three robustness radii on the OOD preference alignment benchmark using the LLaMA-3.2-1B-Instruct model. Figures 2 and 3 show the results obtained by varying one radius at a time while keeping the other robustness components fixed. The results suggest that TPO is generally stable across a broad range of values. The group radius ρ^{ω} has a relatively mild effect, while performance improves as the prompt radius ρ_g^X increases from very small values and then quickly saturates. The preference radius ρ_g^S is more influential: moderate values improve robustness, but overly large values can reduce performance, indicating possible over-regularization. These trends offer practical guidance for robustness-radius selection and motivate future work on task-adaptive tuning strategies.

Effect of Model Family and Scale. To study the effect of model family and scale, we further evaluate TPO on Qwen2.5-3B and LLaMA-3.2-3B. Together with Qwen2.5-0.5B and LLaMA-3.2-1B, this covers two model families, each evaluated at two model sizes. Table 7 reports the detailed results across ID and the four OOD environments, while Table 8 summarizes performance averaged over Group, Prompt, Preference, and Combined. As shown in Table 8, TPO improves both average and worst-group accuracy for all four models, with the largest gains observed on LLaMA-3.2-3B. Within each model family, lower DPO baseline accuracy is associated with larger gains in average accuracy, suggesting a headroom effect. In contrast, worst-group improvements show no consistent trend with model scale. Overall, these results suggest that TPO remains effective across model families, scales, and OOD shift settings.

Downstream Evaluation with Robust DPO Baselines. We further evaluate the downstream generalization of robust preference optimization methods on the Open LLM Leaderboard v2 (Hugging Face, 2024). In addition to standard DPO and TPO, we include two representative robust DPO baselines: WDPO (Xu et al., 2025) and Dr.DPO (Wu et al., 2025). All methods use LLaMA-3.2-1B-Instruct with the same UltraFeedback preference data, SFT initialization, training schedule, and

Table 7: Preference accuracy (%) across model families and scales.

Model	Method	ID		Group		Prompt		Preference		Combined	
		Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg	Worst
Qwen2.5-0.5B	DPO	86.87	77.78	81.80	76.15	83.17	76.98	82.93	78.16	82.60	76.40
	TPO	**86.95**	**85.12**	**84.61**	**81.65**	**85.47**	**82.56**	**85.44**	**82.34**	**85.25**	**79.94**
LLaMA-3.2-1B	DPO	88.75	86.19	88.23	83.44	87.81	85.43	86.91	**81.42**	87.44	82.61
	TPO	**90.57**	**87.55**	**89.11**	**83.50**	**89.14**	**86.70**	**88.25**	81.31	**88.31**	**83.19**
Qwen2.5-3B	DPO	78.60	77.48	76.04	74.53	77.38	76.92	73.97	72.62	76.92	71.99
	TPO	**79.69**	**78.12**	**79.56**	**78.47**	**80.34**	**77.61**	**80.14**	**79.75**	**80.08**	**73.26**
LLaMA-3.2-3B	DPO	85.24	84.04	83.54	80.69	86.20	84.68	82.74	80.00	83.43	80.83
	TPO	**90.55**	**88.56**	**89.82**	**86.92**	**91.05**	**89.12**	**87.36**	**84.31**	**88.66**	**84.72**

Table 8: Average OOD performance and absolute gain of TPO over DPO.

Model	OOD Average			OOD Worst		
	DPO	TPO	Gain	DPO	TPO	Gain
Qwen2.5-0.5B	82.63	85.19	+2.57	76.92	81.62	+4.70
Qwen2.5-3B	76.08	80.03	+3.95	74.02	77.27	+3.26
LLaMA-3.2-1B	87.60	88.70	+1.11	83.23	83.68	+0.45
LLaMA-3.2-3B	83.98	89.22	+5.24	81.55	86.27	+4.72

evaluation protocol. We evaluate the resulting models on MMLU-Pro, BBH, GPQA, MATH-Hard, IFEval, and MuSR. As shown in Table 9, the methods exhibit different strengths across downstream tasks. Dr.DPO achieves the best performance on BBH, GPQA, MATH-Hard, and MuSR, while TPO performs best on MMLU-Pro and IFEval. Overall, the performance differences are relatively small across most benchmarks, with TPO remaining competitive with both standard DPO and the robust DPO baselines. These results suggest that the robustness benefits of TPO are compatible with strong general downstream performance.

Computational Cost. We report the wall-clock training time of LLaMA-3.2-1B-Instruct on the UltraFeedback setting in Figure 4 using 2 NVIDIA H100 80GB GPUs. The time is measured in minutes. Here, TPO$(1,0,0)$, TPO$(1,1,0)$, and TPO$(1,1,1)$ denote enabling group-level robustness only, enabling both group- and prompt-level robustness, and enabling all three robustness components, respectively. The corresponding overheads are approximately $+3.5$, $+13.8$, and $+14.9$ minutes compared with DPO. These results show that TPO introduces additional computation compared with standard DPO, mainly due to group reweighting and the inner adversarial prompt perturbation steps. Nevertheless, the overall overhead remains modest and manageable in our implementation.

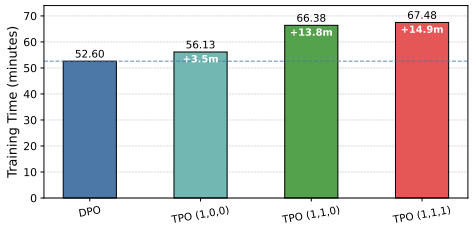


Figure 4: Computational cost on UltraFeedback with LLaMA-3.2-1B-Instruct.

D Limitations

First, our finite-sample convergence guarantee relies on the uniform covariance coverage assumption (Assumption 2), which requires sufficient coverage of pairwise feature differences uniformly over admissible group mixtures and worst-case prompt perturbations. This assumption is standard in preference learning theory, though relaxing it to broader settings remains an interesting theoretical direction.

Second, our prompt-level robustification uses embedding-space perturbations as a practical approximation to shift in the discrete token space, which is a widely adopted strategy in robust NLP but may not capture all forms of prompt-level variation in real-world scenarios, though it has been shown to be effective in our controlled OOD alignment experiments. Further work may explore more robust

Table 9: Downstream evaluation on the Open LLM Leaderboard v2 using LLaMA-3.2-1B-Instruct. The best result is shown in bold and the second-best is underlined.

Method	MMLU-Pro	BBH	GPQA	MATH-Hard	IFEval	MuSR	Avg.
DPO	<u>20.0</u>	35.6	23.2	8.1	51.3	31.6	28.30
WDPO	19.4	<u>36.1</u>	<u>24.1</u>	8.0	47.9	<u>32.5</u>	28.00
Dr.DPO	19.7	**36.4**	**24.3**	**8.6**	<u>52.1</u>	**32.6**	**28.95**
TPO ($\rho_g^S = 0.01$)	**20.5**	35.9	23.7	<u>8.3</u>	**53.0**	31.2	<u>28.77</u>

prompt-level perturbations in the discrete token space, such as adversarial token substitutions or paraphrasing.

Third, TPO introduces additional hyperparameters, including the robustness radii $\{\rho^\omega, \rho_g^X, \rho_g^S\}$. Since the appropriate ranges of these radii may vary across tasks and scenarios, selecting them can require additional effort. Although our sensitivity analysis provides some practical guidance, developing principled selection rules or task-adaptive tuning strategies remains an important direction for future work.

E Impact Statement

This work proposes a hierarchical distributionally robust framework for preference optimization, with the primary goal of improving the reliability of LLM alignment across diverse user populations and deployment environments, which may help reduce performance disparities across underrepresented groups. As a foundational algorithmic contribution, it does not involve model deployment. More robustly aligned models could nonetheless be misused to generate more persuasive harmful content across a broader range of contexts, a risk shared with LLM alignment research more broadly.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our paper's contributions and scope have been clearly described in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Please refer to Appendix D.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

Justification: We have clearly stated all assumptions used in our work.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide dataset construction, model training, hyperparameter search, and evaluation details in Appendices B.1, B.2, and B.3.

   Guidelines:

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

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

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

Answer: [No]

Justification: While the data used in our study is publicly available, we are currently unable to provide open access to the code. However, we have included detailed descriptions of the data access, preprocessing steps, model architecture, and experimental setup in the paper and Appendix. These details should be sufficient for others to reproduce the experimental results.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Please refer to Appendix B.1, B.2, and B.3.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We report mean and standard deviation over three random seeds for the main Emotion benchmark in Table 6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide the details in Appendix B.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We acknowledge the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: See Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This paper has no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We state the assumptions explicitly and provide complete proofs of the theoretical results in the appendix.

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [NA]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [NA]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLMs are the target models in our empirical evaluation, and the paper describes the model backbones, fine-tuning setup, and evaluation protocols.

    Guidelines:

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