Robust Nash Alignment under Preference Uncertainty

Shihab Ahmed
University of Central Florida
shihab.ahmed@ucf.edu

Debamita Ghosh
University of Central Florida
debamita.ghosh@ucf.edu

David Tang
University of Central Florida
david.tang@ucf.edu

Yudan Wang
University of Central Florida
yudan.wang@ucf.edu

Alvaro Velasquez
University of Colorado Boulder
alvaro.velasquez@colorado.edu

Yue Wang
University of Central Florida
yue.wang@ucf.edu

Abstract

Preference-based alignment methods typically optimize against a single preference model, and can therefore be brittle when pairwise preferences are noisy, heterogeneous, or shift after deployment. We propose Robust Nash Alignment, a game-theoretic framework for alignment under uncertain pairwise preferences. The learner maximizes a KL-regularized worst-case payoff against both a competing policy and a preference kernel in a Bernoulli-KL ambiguity set around a nominal estimate. We allow the learner and competitor to use different KL regularization strengths. To address the coupled inner minimization, we introduce a four-player primal-dual proxy involving the leader policy, follower policy, adversarial kernel, and dual variable, and develop a single-loop optimistic mirror descent-ascent algorithm. We prove that the proxy lower-bounds the truncated hard-constrained objective and quantify its approximation error. For fixed truncation, the averaged proxy duality gap is $O(1/T)$. Balancing the truncation with the iteration budget yields an $O(T^{-1/2}) + \rho\gamma_{\min}$ bound on robust suboptimality over the prescribed leader policy class. Experiments on tabular games and LLM alignment illustrate robustness to preference perturbations.

1 Introduction

Aligning large language models (LLMs) with preferences is a central challenge in modern post-training. The dominant paradigm, reinforcement learning from human feedback (RLHF) [9, 41, 33], collects pairwise comparisons, fits a scalar reward model, and optimizes policy against that reward with a proximity constraint to a reference model. More recent direct alignment methods, including Direct Preference Optimization (DPO), SLiC-HF, Contrastive Preference Learning (CPL), and Soft Preference Optimization (SPO) [35, 55, 15, 40], bypass the explicit reward model by optimizing preference-derived surrogate losses directly. Despite their practical success, all of these methods share a common structural assumption: they model preferences through a latent scalar reward function, typically via the Bradley-Terry (BT) [45, 4] or Thurstone framework. This assumption is, however, restrictive. Preferences may be context-dependent, cyclic, or heterogeneous across annotators, and cannot always be represented by a single scalar reward [3, 29, 54, 52, 8, 14].

A growing line of work addresses this limitation by modeling preferences directly through pairwise comparison probabilities and framing alignment as a two-player zero-sum game. In particular, Nash

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

Learning from Human Feedback (NLHF) and its extensions [29, 54, 52, 5] seek a policy that is hard to beat in the induced preference game, rather than one that maximizes a learned scalar reward. This general-preference perspective resolves the modeling mismatch: it can represent non-transitive, cyclic, and population-heterogeneous preferences that no scalar reward can capture. However, it introduces a different and equally important challenge that has been largely overlooked: *the preference kernel itself is uncertain.*

The deployment pairwise preference probability $P(y_i \succ y_j \mid x)$ is generally unknown and may change over time. In practice, it must be estimated from finite, noisy, and potentially biased comparisons, often aggregated across heterogeneous annotator populations and changing deployment environments [9, 41, 33, 43]. Existing general-preference methods, including NLHF, treat the learned preference kernel as a fixed ground truth and optimize against this single point estimate. This can be brittle: small errors in preference estimation, misspecification of the comparison model, or deployment-time shifts in who provides feedback may materially change which responses the policy favors [36, 38, 8, 14]. We refer to this vulnerability as the *fragile consensus* problem—a policy that appears optimal under the nominal preference data may concentrate on responses whose advantages rest on slim margins that are statistically indistinguishable from noise, and a small perturbation to the preference kernel can reverse the dominance ordering entirely.

To address this challenge, we propose to treat the preference kernel as fundamentally uncertain and to optimize for the worst case over a set of plausible kernels. Inspired by the literature on distributionally robust optimization [36, 20, 31], where ambiguity sets, divergence balls, and worst-case guarantees are standard tools for hedging estimation error and distribution shift, we formulate *alignment under uncertain preferences* as a robust two-player game over a pairwise preference kernel. At each prompt, the main player seeks to maximize a Kullback-Leibler (KL)-regularized win rate, while an adversarial player simultaneously chooses the hardest competitor *and* the worst-case preference kernel from an ambiguity set centered at the nominal estimate. The resulting objective provides a lower bound on the regularized payoff under every preference kernel in the ambiguity set.

Our contributions are summarized as follows.

Robust preference-game formulation. We formulate alignment under preference uncertainty as a robust game over a pairwise preference kernel with a Bernoulli-KL ambiguity set. Unlike standard NLHF, which optimizes against a single nominal kernel, our objective maximizes the worst-case regularized game value over kernels within a prescribed divergence ball. The leader and follower may use different KL strengths. The robust value lower-bounds the regularized payoff against every admissible kernel and competing policy.

Four-player algorithm with convergence guarantees. To avoid the computational burden of solving the hard-constrained distributionally robust problem directly, we pass to its Lagrangian as a proxy and derive a four-player primal-dual formulation involving the leader policy, follower policy, adversarial preference kernel, and dual variable. We quantify the gap between the proxy and the hard-constrained robust objective and prove an $O(1/T)$ averaged proxy duality-gap bound for fixed truncation. With truncation of order $T^{-1/2}$, this yields an $O(T^{-1/2}) + \rho\gamma_{\min}$ robust suboptimality bound over the floored leader simplex.

Experimental evidence. We evaluate practical variants on controlled tabular games, a best-of-m reranking task on IMDb continuations with heterogeneous annotator groups, and a summarization task on TL;DR. The reported results compare worst-case values and preference outcomes under perturbed kernels and external judges. The experimental procedures and their relation to the tabular theory are discussed in Section 5.

Related work. Distributionally robust optimization hedges estimation error and shift with ambiguity sets [38, 20, 31], robust alignment methods robustify reward-based or DPO-style objectives [6, 24], and robust (Markov) game theory treats payoff or transition uncertainty inside equilibrium concepts [1, 22, 53]. We instead robustify the pairwise kernel itself inside a general-preference game, with a coupled, non-rectangular budget and an asymmetric leader–follower objective (Appendix A).

2 Preliminaries

Notation. Let $\mathcal{X}$ denote the prompt space, and let D_0 be the deployment-time prompt distribution. For each prompt $x \in \mathcal{X}$, let $Y(x) = \{y_1, \ldots, y_m\}$ be the finite response set. A tabular policy assigns

a distribution $\pi(\cdot \mid x) \in \Delta(Y(x))$ to each prompt x. over responses for prompt x. Given x, a response y is sampled from $\pi(\cdot \mid x)$. We also fix a reference policy $\pi_{\mathrm{ref}}(\cdot \mid x)$ with full support, which can be derived from Supervised Fine-Tuning (SFT), and a KL regularization parameter $\tau > 0$.

Reward-based alignment. A standard approach to preference-based alignment assumes that pairwise human preferences are generated from an underlying scalar reward function $R^\star(x, y)$. Under the classical Bradley-Terry / logistic preference model [4, 9, 33], the probability that response y is preferred to y' under prompt x is

$$\mathbb{P}(y \succ y' \mid x) = \frac{\exp(R^\star(x, y))}{\exp(R^\star(x, y)) + \exp(R^\star(x, y'))}. \tag{1}$$

Given a dataset with pairwise responses and preferences, one can first learn a reward model R through regression, and maximize the classical KL-regularized RLHF objective

$$J_{\mathrm{RB}}(\pi; R) := \mathbb{E}_{x\sim D_0}\left[\mathbb{E}_{y\sim\pi(\cdot|x)}[R(x, y)] - \tau D_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big)\right]. \tag{2}$$

The reward-based alignment problem is then $\pi^\star_R \in \arg\max_{\pi\in\Pi} J_{\mathrm{RB}}(\pi; R)$, where Π is the policy class of interest. The KL term discourages the learned policy from drifting arbitrarily far from the reference policy [41, 33, 35]. For unrestricted tabular policies, the maximizer has the closed form

$$\pi^\star_R(y \mid x) \propto \pi_{\mathrm{ref}}(y \mid x) \exp\left(\frac{1}{\tau} R(x, y)\right), \tag{3}$$

which underlies the derivation of the direct preference optimization method [35].

General preference alignment. A limitation of (1) is that it assumes all pairwise preferences are induced by a single scalar reward function. More generally, one can directly treat pairwise preference probabilities as primitive objects, without assuming the latent reward [29, 52, 54].

Definition 1. A general preference oracle is a mapping $P : \mathcal{X} \times Y \times Y \to [0, 1]$ such that, for every prompt $x \in \mathcal{X}$ and responses $y, y' \in Y(x)$, there exists a preference $z \sim \mathrm{Ber}\big(P(y \succ y' \mid x)\big)$, where $z = 1$ indicates that y is preferred to y', and $z = 0$ indicates the opposite.

Assumption 1 (Antisymmetric preference model)**.** *For any fixed prompt x, define $P_x(y \succ y') := P(y \succ y' \mid x)$. We assume that the preference probabilities are antisymmetric, i.e.,*

$$P_x(y_i \succ y_j) + P_x(y_j \succ y_i) = 1, \qquad \forall\, y_i \neq y_j, \qquad P_x(y_i \succ y_i) = \frac{1}{2}. \tag{4}$$

Given two policies $\pi, \nu \in \Pi$, Nash learning from human feedback (NLHF) considers the KL-regularized game payoff under the general preference kernel P:

$$\begin{aligned} J(\pi, \nu; P) := \mathbb{E}_{x\sim D_0}\Big[&\mathbb{E}_{y\sim\pi(\cdot|x),\ y'\sim\nu(\cdot|x)}\big[P_x(y \succ y')\big] - \tau D_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big) \\ &+ \tau D_{\mathrm{KL}}\big(\nu(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big)\Big]. \end{aligned} \tag{5}$$

Here, π is the max-player and ν is the min-player. The interpretation is that both players seek a policy that is consistently preferred to any competitor, while both players are regularized toward the same reference policy. The nominal general-preference alignment problem is therefore formulated as a regularized minimax game

$$\pi^\star \in \arg\max_{\pi\in\Pi} \min_{\nu\in\Pi} J(\pi, \nu; P), \qquad \nu^\star \in \arg\min_{\nu\in\Pi} \max_{\pi\in\Pi} J(\pi, \nu; P). \tag{6}$$

For unrestricted tabular policies at a fixed prompt, the symmetric convex–concave game has a unique symmetric Nash equilibrium $(\pi^\star, \pi^\star)$ [29]. NLHF seeks this policy as the alignment policy.

3 Robust Game Formulation for Uncertain Preference Alignment

In this section, we introduce our formulation of a robust game to address uncertainty in preferences. For notational simplicity, we fix a prompt and suppress the prompt index. The global prompt-averaged objective is obtained by taking expectations under a prompt-rectangular ambiguity model.

Let $Y = \{1, \dots, m\}$, with $m \geq 2$, be the response set, let $d = \binom{m}{2}$, and set $E := \{(i,j) : 1 \leq i < j \leq m\}$. For $p = (p_{ij})_{(i,j)\in E} \in [0,1]^d$, define the preference kernel satisfying Assumption 1 by

$$P_p(i \succ j) = \begin{cases} p_{ij}, & i < j, \\ 1 - p_{ji}, & i > j, \\ 0.5, & i = j. \end{cases}$$

Here $i < j$ is the ordinary index order: each unordered pair is listed once, and the reverse comparison follows by antisymmetry.

We model the potential uncertainty in the preference kernel through the Bernoulli-KL divergence. Namely, centered at a nominal kernel $p^\star \in (0,1)^d$, the Bernoulli-KL divergence is defined as

$$D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) := \sum_{(i,j)\in E} \left[p_{ij} \log \frac{p_{ij}}{p^\star_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - p^\star_{ij}} \right], \tag{7}$$

measuring the difference between the two kernels. We then set the ambiguity set as

$$\mathcal{P} := \{p \in [0,1]^d : D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) \leq \rho\}, \tag{8}$$

where $\rho \geq 0$ is the uncertainty radius. Thus the ambiguity set contains kernels whose Bernoulli-KL divergence from the nominal kernel is at most ρ.

Remark 1. The ambiguity set $\mathcal{P}$ captures preference kernels that are close to a nominal estimate $p^\star$, and thus reflects uncertainty arising from finite, noisy, or heterogeneous preference data. The radius ρ is generally chosen by domain experts or users, reflecting the users' consideration level of uncertainty. For instance, in offline preference-based learning, one may estimate $p^\star$ from pairwise comparisons and calibrate ρ under an appropriate sampling model so that $\mathcal{P}$ covers the deployment kernel with a prescribed confidence level. When preferences are aggregated across different users or annotator populations, $\mathcal{P}$ can also model variability in pairwise comparisons.

Unlike uncertainty sets that decouple across components [48, 20], $\mathcal{P}$ couples all pairwise comparisons through one divergence budget. The coupled minimization over the competing policy and kernel is the optimization challenge studied below.

Objective. With the ambiguity set defined in (8), we propose our robust formulation as

$$\max_{\pi \in \Delta_m} \min_{\nu \in \Delta_m} \min_{p \in \mathcal{P}} J(\pi, \nu; p), \tag{9}$$

We use positive KL strengths τ_π and τ_ν for the leader and follower, respectively. They may differ, allowing the two policies to be anchored to the reference with different strengths. The payoff under kernel p is

$$J(\pi, \nu; p) := \sum_{i=1}^{m} \sum_{j=1}^{m} \pi_i \nu_j \, P_p(i \succ j) - \tau_\pi D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}}) + \tau_\nu D_{\mathrm{KL}}(\nu \| \pi_{\mathrm{ref}}). \tag{10}$$

The adversary chooses both a competing policy and an admissible preference kernel. Let $V_{\mathrm{full}}(\rho)$ denote the optimal value of (9). If $\pi^\star$ solves (9), then, for every $p \in \mathcal{P}$ and $\nu \in \Delta_m$,

$$\begin{aligned} &J(\pi^\star, \nu; p) \geq V_{\mathrm{full}}(\rho), \\ &P_p(\pi^\star \succ \nu) \geq V_{\mathrm{full}}(\rho) + \tau_\pi D_{\mathrm{KL}}(\pi^\star \| \pi_{\mathrm{ref}}) - \tau_\nu D_{\mathrm{KL}}(\nu \| \pi_{\mathrm{ref}}). \end{aligned} \tag{11}$$

Here $P_p(\pi \succ \nu) := \sum_{i,j} \pi_i \nu_j P_p(i \succ j)$. The optimized certificate concerns the regularized payoff; the displayed KL terms give the corresponding raw win-rate bound. Therefore, the maximizer $\pi^\star$ of our robust value (9) provides a baseline performance of uncertain alignment, thus improving the robustness and reliability against preference uncertainty.

4 Four-Player Proxy Game

In this section, we aim to design a concrete algorithm to solve the robust game (9).

For standard non-robust NLHF, the kernel is fixed and the two policy blocks form a convex–concave game. In our formulation, even for a fixed leader π, the inner value $g_H(\pi; \rho) :=$

$\min_{\nu\in\Delta_m,\, p\in\mathcal{P}} J(\pi,\nu;p)$ couples ν and p through the win-rate term. If one first fixes π, p and eliminates the adversarial policy $\nu^\star$ using the Gibbs response, the reduced objective becomes

$$\min_{p\in\mathcal{P}}\Big\{-\tau_\pi D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}})-\tau_\nu\log\sum_{j=1}^{m}\pi_{\mathrm{ref}}(j)\exp\big(-\frac{b(j;\pi,p)}{\tau_\nu}\big)\Big\},\tag{12}$$

where $b(j;\pi,p)=\sum_i \pi_i P_p(i\succ j)$ is affine in p. The negative log-sum-exp makes the reduced objective concave in p, so this is generally a nonconvex minimization problem. Concave minimization is difficult in general [18]; this observation alone does not establish NP-hardness for the structured family considered here. For fixed (π,ν), however, the kernel subproblem is convex: its objective is affine in p and its Bernoulli-KL constraint is convex. The difficulty comes from optimizing (ν,p) jointly. Their bilinear interaction prevents a direct application of the convex–concave saddle-point analysis used for nominal NLHF.

Due to these challenges, we develop a proxy game formulation that is easier to solve and develop detailed studies on the connection between the proxy and the original robust game.

For regularity, we first assume the main player policy is constrained to the floor-truncated simplex $\Pi_\epsilon:=\{\pi\in\Delta_m:\ \pi_i\ge\epsilon\,\pi_{\mathrm{ref},i}\ \ \forall i\}$, where $0<\epsilon\le 1$. The convergence guarantees below compare against this prescribed leader class. For each $\pi\in\Pi_\epsilon$, denote the hard-constrained inner value by

$$g_H(\pi;\rho):=\min_{\nu\in\Delta_m}\min_{p\in\mathcal{P}} J(\pi,\nu;p),\tag{13}$$

and the corresponding optimal robust value as

$$V(\rho):=\max_{\pi\in\Pi_\epsilon} g_H(\pi;\rho).\tag{14}$$

We further define an interior box $P_\delta:=[\delta,1-\delta]^d$, where $\delta\in(0,1/2)$, and consider the truncated hard objective:

$$g_H^\delta(\pi;\rho):=\min_{\nu\in\Delta_m}\min_{p\in P_\delta\cap\mathcal{P}} J(\pi,\nu;p),\tag{15}$$

and denote its optimal value as

$$V^\delta(\rho):=\max_{\pi\in\Pi_\epsilon} g_H^\delta(\pi;\rho).\tag{16}$$

We first show that the truncation error is explicit and uniformly controlled.

Theorem 1 (Uniform approximation by the truncated hard objective)**.** *Assume $p^\star\in P_\delta$. Then, for every $\pi\in\Pi_\epsilon$, $0\le g_H^\delta(\pi;\rho)-g_H(\pi;\rho)\le\delta$ and $0\le V^\delta(\rho)-V(\rho)\le\delta$. Moreover, if $\pi_\delta^\star\in\arg\max_{\pi\in\Pi_\epsilon} g_H^\delta(\pi;\rho)$, then $g_H(\pi_\delta^\star;\rho)\ge V(\rho)-\delta$.*

This theorem shows that the error introduced by the regularity relaxation is bounded by δ. Thus, one can choose δ small enough to ensure regularity and approximate the original solution arbitrarily. Therefore, we will focus on solving this problem, $\arg\max_{\pi\in\Pi_\epsilon} g_H^\delta(\pi;\rho)$. Since g_H^δ is a hard-constrained problem, which can be inefficient to solve directly, we introduce a Lagrangian dual variable and consider the Lagrangian function. Specifically, with an additional dual variable γ, we define and consider the four-player payoff as

$$F(\pi,\gamma,\nu,p):=J(\pi,\nu;p)+\gamma\big(D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star)-\rho\big).\tag{17}$$

Moreover, fix a compact positive interval $\Gamma:=[\gamma_{\min},\gamma_{\max}]\subset(0,\infty)$ and $\mathcal{N}_\beta:=\{\nu\in\Delta_m:\ \nu_j\ge\beta\,\pi_{\mathrm{ref},j}\ \ \forall j\}$ for $\beta\in(0,1)$, we define the proxy objective as

$$g_{\mathrm{proxy}}^\delta(\pi;\rho):=\max_{\gamma\in\Gamma}\min_{\nu\in\mathcal{N}_\beta,\, p\in P_\delta} F(\pi,\gamma,\nu,p),\tag{18}$$[1]

and its optimal value as

$$V_{\mathrm{proxy}}^\delta(\rho):=\max_{\pi\in\Pi_\epsilon} g_{\mathrm{proxy}}^\delta(\pi;\rho).\tag{19}$$

We then study the proxy's connection to the original problem. Firstly, we introduce a regularity assumption.

[1] Under Assumption 2-(2), restricting the follower to $\mathcal{N}_\beta$ leaves each fixed-kernel best-response value unchanged; see Appendix B.3.

Assumption 2. *We assume the following regularization conditions hold: (1) $p^\star \in (\delta, 1-\delta)^d$; (2) $0 < \beta \le e^{-1/\tau_\nu}$; (3) $\rho > 0$ and $\Gamma = [\gamma_{\min}, \gamma_{\max}] \subset (0, \infty)$ with $\gamma_{\max} \ge 1/\rho$; (4) $\tau_\nu \gamma_{\min} \ge \frac{1}{4}$.*

Parts (1)–(3) specify the nominal-kernel interior, follower floor, and multiplier range. Part (4) ensures joint convexity in the follower and kernel. The leader coefficient τ_π need only be positive.

We then derive the following results for the connections.

Theorem 2. *Under (1)-(3) of Assumption 2, for every $\pi \in \Pi_\epsilon$, the proxy value always lower bounds the hard-constrained value:*

$$g_H^\delta(\pi;\rho) = \min_{\nu \in \mathcal{N}_\beta} \max_{\gamma \in [0,\gamma_{\max}]} \min_{p \in P_\delta} F(\pi,\gamma,\nu,p) \ge g_{\text{proxy}}^\delta(\pi;\rho). \tag{20}$$

If we additionally assume (4) of Assumption 2, then for every $\pi \in \Pi_\epsilon$,

$$0 \le g_H^\delta(\pi;\rho) - g_{\text{proxy}}^\delta(\pi;\rho) \le \rho\gamma_{\min}, \text{ and } 0 \le V^\delta(\rho) - V_{\text{proxy}}^\delta(\rho) \le \rho\gamma_{\min}. \tag{21}$$

The theorem relates the proxy value g_{proxy}^δ to the truncated hard value g_H^δ. Under parts (1)–(3), the proxy is a lower bound. Part (4) bounds the difference by $\rho\gamma_{\min}$; combining this with Theorem 1 gives $|g_H(\pi;\rho) - g_{\text{proxy}}^\delta(\pi;\rho)| \le \delta + \rho\gamma_{\min}$. The approximation term remains in the final robust guarantee. Its trade-off with curvature is governed by $\tau_\nu\gamma_{\min} \ge 1/4$.

The proxy exchanges the order of $\max_\gamma$ and $\min_\nu$ in (20), giving the grouped problem $\max_{\pi,\gamma} \min_{\nu,p} F$. Under Assumption 2, F is concave in (π,γ) and convex in (ν,p), so compactness and the minimax theorem give a saddle point. The next section develops a single-loop algorithm for this problem.

5 Algorithm Design and Convergence Analysis

We treat each block as a player and design a four-player gradient method for the proxy. Under Assumption 2, the grouped proxy has a monotone saddle operator. Since simultaneous gradient methods can still cycle in convex–concave games, we use optimistic mirror descent-ascent, building on optimistic and extragradient methods [7, 37, 42, 27, 28, 47]. Specifically, we aim to solve the proxy $\max - \min$ problem

$$\max_{x \in X} \min_{y \in Y} F(x,y), \quad x = (\pi,\gamma) \in X := \Pi_\epsilon \times \Gamma,\ y = (\nu,p) \in Y := \mathcal{N}_\beta \times P_\delta,$$

by a single-loop optimistic mirror descent/ascent method.

We use entropy geometry for the simplex blocks, Euclidean geometry for γ, and binary entropy geometry for p. The associated Bregman divergences are: $D_\pi(\pi,\pi') = D_{\text{KL}}(\pi\|\pi')$, $D_\gamma(\gamma,\gamma') = \frac{1}{2}(\gamma-\gamma')^2$, $D_\nu(\nu,\nu') = D_{\text{KL}}(\nu\|\nu')$, and $D_p(p,p') = D_{\text{KL}}^{\text{Bern}}(p\|p')$. At iteration t, let

$$z_t = (\pi_t, \gamma_t, \nu_t, p_t) \in X \times Y, \qquad g_t^\bullet := \nabla_\bullet F(z_t),$$

and define $\widehat{g}_t^\bullet := 2g_t^\bullet - g_{t-1}^\bullet$, with $\bullet \in \{\pi,\gamma,\nu,p\}$. The block mirror updates are given in Appendix B.1, Eqs. (26)–(29), and summarized in Algorithm 1. Since the mirror maps are separable across blocks, these coordinate-wise updates are exactly optimistic mirror descent/ascent on the grouped saddle problem $\max_{x\in X} \min_{y\in Y} F(x,y)$.

We use the product norm

$$\|z - z'\|^2 := \|\pi - \pi'\|_1^2 + |\gamma - \gamma'|^2 + \|\nu - \nu'\|_1^2 + \|p - p'\|_2^2$$

and its dual norm. On $Z := X \times Y$, define the signed saddle gradient

$$G(z) := (-\nabla_\pi F, -\partial_\gamma F, \nabla_\nu F, \nabla_p F).$$

$$H_\delta := \sup_{z \in Z} \|G(z)\|_*, \qquad R := \sup_{z,z' \in Z} \|z - z'\|.$$

Both constants are finite because the domains are compact and all coordinates in logarithms are bounded away from zero.

Algorithm 1 Four-player optimistic gradient descent/ascent (OGDA)

Require: step size $\eta > 0$, iteration budget T, initial point $z_1 = (\pi_1, \gamma_1, \nu_1, p_1) \in X \times Y$

Compute $g_1^\bullet = \nabla_\bullet F(z_1)$ for all blocks, and set $g_0^\bullet \leftarrow g_1^\bullet$
for $t = 1, \dots, T$ **do**
  Form optimistic gradients $\widehat{g}_t^\bullet = 2g_t^\bullet - g_{t-1}^\bullet$
  Update $(\pi_{t+1}, \gamma_{t+1}, \nu_{t+1}, p_{t+1})$ by (26)-(29)
  if $t < T$ **then**
    Compute $g_{t+1}^\bullet = \nabla_\bullet F(z_{t+1})$ for all blocks
  end if
end for
Return the averaged iterate $\bar{z}_T = (\bar{\pi}_T, \bar{\gamma}_T, \bar{\nu}_T, \bar{p}_T) := \frac{1}{T}\sum_{t=1}^{T}(\pi_t, \gamma_t, \nu_t, p_t)$.

Theorem 3 (Convergence and robust guarantee of four-player OGDA)**.** *Assume Assumption 2, and run Algorithm 1 with $\eta = 1/(8L_\delta)$, where L_δ is the constant in Lemma 3. Define*

$$D_X^{\mathrm{diam}} := \sup_{(\pi,\gamma)\in X} \left\{ D_{\mathrm{KL}}(\pi\|\pi_1) + \tfrac{1}{2}(\gamma - \gamma_1)^2 \right\},$$
$$D_Y^{\mathrm{diam}} := \sup_{(\nu,p)\in Y} \left\{ D_{\mathrm{KL}}(\nu\|\nu_1) + D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p_1) \right\},$$
$$C_\delta := 8L_\delta(D_X^{\mathrm{diam}} + D_Y^{\mathrm{diam}}) + H_\delta R + \frac{H_\delta^2}{16L_\delta}.$$

The averaged iterate satisfies

$$\mathrm{Gap}(\bar{z}_T) := \max_{x'\in X} F(x', \bar{y}_T) - \min_{y'\in Y} F(\bar{x}_T, y') \le \frac{C_\delta}{T}, \tag{22}$$

and its leader policy satisfies

$$V(\rho) - g_H(\bar{\pi}_T; \rho) \le \frac{C_\delta}{T} + \rho\gamma_{\min} + \delta. \tag{23}$$

Fix all parameters except δ, initialize $\pi_1 = \nu_1 = \pi_{\mathrm{ref}}$ and $p_1 = p^\star$, and fix $\gamma_1 \in \Gamma$. For $\delta_T = \min\{cT^{-1/2}, \delta_0\}$, where $c > 0$ and $0 < \delta_0 < \min_e\{p_e^\star, 1 - p_e^\star\}$ are fixed, the robust suboptimality over Π_ϵ is $O(T^{-1/2}) + \rho\gamma_{\min}$.

Thus the averaged iterates have a vanishing proxy saddle gap, and the averaged leader has the stated hard-objective approximation guarantee over Π_ϵ. Appendix C compares this analysis with nominal NLHF.

Cost. Each iteration uses $O(m^2)$ arithmetic for the gradients and $O(m^2)$ memory; the floored-simplex projections have an additional numerical-precision cost. With exact updates, $T = \lceil C_\delta/\varepsilon_{\mathrm{acc}} \rceil$ iterations suffice for proxy gap at most $\varepsilon_{\mathrm{acc}}$. A smaller δ reduces truncation error but increases the convergence constant. The multiplier range and the follower coefficient τ_ν govern the approximation and curvature trade-offs; Appendix B.5 gives their dependence.

From tabular NLHF to LLM fine-tuning. Theorems 1–3 concern finite response sets, exact block gradients, and a policy-independent sum of Bernoulli divergences. For free-form generation, Section 6.3 uses the sampled procedure in Appendix F.1, following the policy-gradient implementation of nominal Nash learning [29]. The nominal baseline uses a geometric-mixture opponent, while the robust procedure trains a follower and an antisymmetric kernel head and updates the multiplier using sampled budget slack. The practical divergence is averaged over sampled response pairs, so its weights depend on the current policies. Together with neural parameterization, alternating adversary steps, and the absence of optimistic updates and iterate averaging, this makes the procedure a training heuristic outside the tabular convergence theorem. The reported experiments retain the shared KL coefficient $\tau_\pi = \tau_\nu = \tau$; they do not evaluate unequal coefficients.

6 Experiments

We then evaluate our approach across three settings of increasing scale: (i) a synthetic tabular game that isolates the fragile-consensus phenomenon and has a known optimal robust value, run with robust

training that solves the inner problem exactly, (ii) a best-of-m reranking task on the IMDb dataset with heterogeneous annotator groups and explicitly perturbed kernels, run with a nested variant of OGDA (Appendix E.2), and (iii) a summarization task with LLM-generated responses and external LLM judgments, run with the sampled analogue of Section 5. None of them verifies the guarantees of Section 5.

6.1 Tabular Game

We first consider a collection of independently generated tabular prompts, each with a finite response set of size $m = 3$. For each prompt x, we construct a *fragile-consensus* preference kernel by drawing a slim margin $s_x \sim \mathrm{Unif}[0.015, 0.040]$ and a solid margin $S_x \sim \mathrm{Unif}[0.25, 0.40]$, and setting

$$P_x^\star \;=\; \begin{pmatrix} 1/2 & 1/2 + s_x & 1/2 - S_x \\ 1/2 - s_x & 1/2 & 1/2 + S_x \\ 1/2 + S_x & 1/2 - S_x & 1/2 \end{pmatrix}. \tag{24}$$

We then apply a random permutation of the response labels so that the identity of the fragile action varies across prompts. This construction produces a cyclic preference game with one easily reversible edge (KL cost $\approx 2s_x^2 \le 0.003$) and two expensive-to-reverse edges (KL cost $\approx 2S_x^2/(S_x(1-S_x)) \ge 0.5$), which is precisely the regime in which robustness matters most: the adversary can flip the cheap edge within a small budget ρ, reversing the dominance ordering, while the expensive edges are untouchable. We use a uniform reference policy and set $\tau = 0.05$.

We compare: (i) **NashMD**, the nominal two-player baseline operating on the fixed kernel $P_x^\star$ (optimistic mirror updates of the two policies, with the kernel frozen); (ii) **Four-player robust training** with training radii $\rho \in \{0.10, 0.20, 0.40\}$; and (iii) the **optimal robust value** $V(\rho)$, computed by golden-section search over the leader policy (g_H is concave in π). Both start from $\pi_1 = \pi_{\mathrm{ref}}$ (NashMD also from $\nu_1 = \pi_{\mathrm{ref}}$) and run for 2000 iterations; the step size is 0.30 for NashMD and 0.05 for robust training. Robust training is not covered by the guarantees of Section 5. We evaluate $g_H(\bar\pi_k; \rho)$, the worst-case value over the ball of radius ρ, at the averaged policy $\bar\pi_k = \frac{1}{k}\sum_{t\le k} \pi_t$.

Figure 1(a) shows $g_H(\bar\pi_k; 0.2)$ over the iterations. NashMD levels off about 0.032 below the optimal robust value $V(0.2)$, confirming that the nominal Nash policy is suboptimal for the robust objective. The robust policy at $\rho = 0.2$ gets close to $V(0.2)$ (mean gap 0.0014 at iteration 2000). At $\rho = 0.1$ (under-budgeted), the policy ends at a lower value, since it optimizes for a smaller perturbation set. At $\rho = 0.4$ (over-budgeted), the policy is more conservative and also ends below the $\rho = 0.2$ value. Figure 1(b) shows the paired per-prompt difference to NashMD of the final policies, with paired-bootstrap 95% intervals over the 60 prompts. The advantage does not simply grow with ρ_{eval}: it peaks somewhat above the training radius (at 0.3, 0.3 and 0.45 for $\rho_{\mathrm{train}} = 0.1, 0.2, 0.4$) and then decreases slightly, and the policy trained at $\rho_{\mathrm{train}} = 0.4$ is worse than NashMD on average at $\rho_{\mathrm{eval}} = 0.05$ and 0.10 (better on 0% and 25% of the prompts), as is the one trained at $\rho_{\mathrm{train}} = 0.2$ at $\rho_{\mathrm{eval}} = 0.05$. At the matched radius $\rho_{\mathrm{eval}} = \rho_{\mathrm{train}}$ the robust policy improves g_H on all 60 prompts, by 0.0160, 0.0309 and 0.0490 for $\rho_{\mathrm{train}} = 0.1, 0.2, 0.4$ (half-widths of the intervals at most 0.0016), closing 91%, 96% and 97% of the gap $V - g_H(\mathrm{NashMD})$, and it improves g_H on all 60 prompts at every larger evaluation radius tested. This further confirms the enhanced robustness of our approach against preference uncertainty.

6.2 Best-of-m IMDb Continuation Task

We next evaluate the method on a best-of-m reranking task, where candidate responses are generated by a pretrained LLM and preference kernels are constructed from heterogeneous annotator groups.

For each of $N = 8000$ prompts from the IMDb review dataset [23], we pre-generate $m = 8$ candidate continuations using `google/flan-t5-large` [11]. We generate the preferences based on three annotator groups: **Comprehensive/analytical** (p_1) prefers detailed, well-structured responses; **Efficient/concise** (p_2) prefers short, direct responses; **Balanced/nuanced** (p_3) prefers hedged, multi-perspective responses (scoring rules in Appendix E.1). We set the nominal kernel as the uniform mixture of the three preferences: $P_{p^\star}(a \succ b) = (\sum_{i=1}^3 P_i(a \succ b))/3$, and construct the uncertainty set centered at it as $\mathcal{P}_{\rho_{\mathrm{eval}}} \triangleq \{(w_1, w_2, w_3) \in \Delta_3 : D_{\mathrm{KL}}((w_1, w_2, w_3)||(1/3, 1/3, 1/3)) \le \rho_{\mathrm{eval}}\}$ to model preference uncertainty.

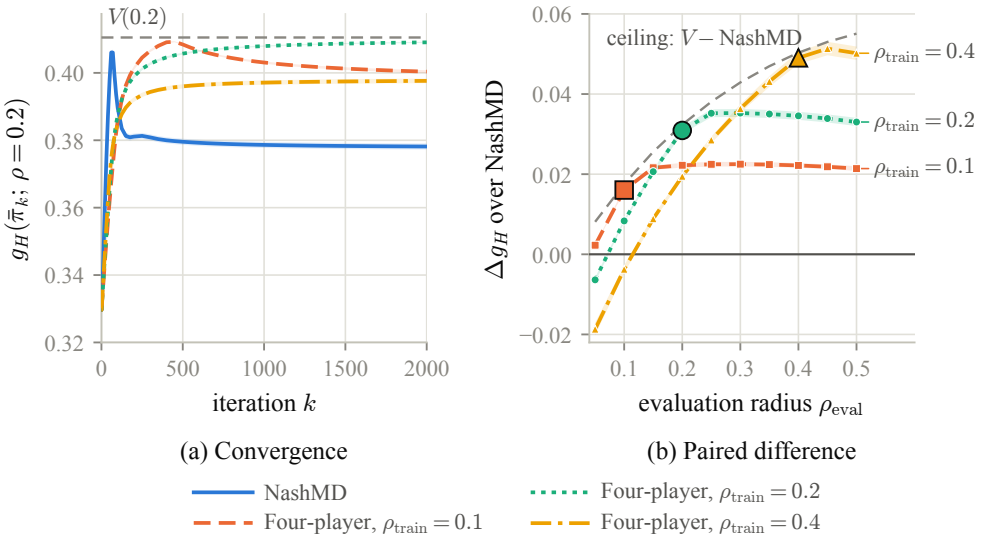


Figure 1: Tabular games. (a) $g_H(\bar{\pi}_k; 0.2)$ of NashMD and of robust training with $\rho_{\text{train}} \in \{0.1, 0.2, 0.4\}$ (running averages; mean and 95% bootstrap interval over prompts) against $V(0.2)$. (b) Paired difference to NashMD of the final averaged policies versus the evaluation radius (95% paired-bootstrap intervals); large markers: matched radius; dashed: ceiling $V(\rho_{\text{eval}}) - g_H(\text{NashMD})$.

We then train and compare the robust policy π_R (trained at radius ρ_{train}) against the NashMD baseline π_N by their pairwise *win rate* under perturbed kernels. For each evaluation radius ρ_{eval}, we sample $M = 600$ kernels $\{P_j\}_{j=1}^{M}$ uniformly from $\mathcal{P}_{\rho_{\text{eval}}}$ and compute the win rate:

$$\text{WinRate}(\pi_R \text{ vs } \pi_N;\, \rho_{\text{eval}}) \;:=\; \big|\{j \in [M] : P_{p_j}(\pi_R \succ \pi_N) \geq 0.5\}\big|/M, \tag{25}$$

where $P_{p_j}(\pi_R \succ \pi_N) = \sum_{i,k} \pi_R(i)\, \pi_N(k)\, p_j(i \succ k)$ is the probability that a response drawn from π_R is preferred over one drawn from π_N under kernel p_j. A win rate above 50% means the robust policy is preferred to NashMD under the majority of plausible kernels. Table 3 (Appendix E.4) reports the win rate of the robust policy against NashMD, averaged over all 8000 prompts, for training radii $\rho_{\text{train}} \in \{0.01, 0.02, 0.05, 0.10, 0.15\}$ and evaluation radii $\rho_{\text{eval}} \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$.

All 25 entries of Table 3 are at least 0.766, also far above the training radius; within a column the win rate decreases as ρ_{train} grows (0.986 to 0.934 at $\rho_{\text{eval}} = 0.05$). In plain terms, a small training radius already protects against kernels far from the nominal one (Table 3, $\tau = 0.01$: 0.986 to 0.844 as ρ_{eval} grows from 0.05 to 0.50), and hedging harder buys nothing on these kernels. The gain depends on the KL temperature (Appendix E.4) and is largest at $\tau = 0.05$ (above 97% for $\rho_{\text{eval}} \leq 0.30$): at very small τ NashMD is already nearly deterministic, at large τ both policies are nearly uniform.

6.3 Summarization Task with LLMs

We further evaluate our performance on the TL;DR summarization task [41]. The nominal preference kernel is a non-Bradley-Terry (non-BT) pairwise judge, $P_\phi^{\text{pair}}(y_a \succ y_b \mid x) = \sigma\big(L^{\text{pair}}(x, y_a, y_b)\big)$, instantiated with PairRM [21]: since each candidate's score is conditioned on its opponent, the kernel is not constrained to admit any scalar reward decomposition. On real model outputs, the closest Bradley-Terry fit to this kernel leaves a mean residual of 0.384 logits per edge (3.10 percentage points), with 1.4% of triples outright cyclic, confirming the nominal kernel used here is measurably non-BT (construction and diagnostic in Appendix F.7). To evaluate robustness under preference uncertainty, we use three external LLMs from different model families: `DeepSeek-R1-Distill-Qwen-32B` [12], `Gemma-4-31B` [44], and `Nemotron-3-Nano-Omni` [32], as judgment at evaluation time. Since each external judge will induce its own pairwise preference (from their own reasoning) that differs from the nominal one, these judgments simulate a deployment scenario where the user population's preferences diverge from the training-time annotation distribution.

The results are presented in Table 1. Each entry is the share of evaluation prompts on which the judge prefers the robust policy's summary to NashMD's; the judges never entered training, so a win means the policy holds up under a preference it was not trained on. Most external judges (five of six configurations) yield a win rate above 0.50 over NashMD, confirming that the robust policy remains competitive under this non-BT nominal kernel even though, unlike prior work, our framework does not require a Bradley-Terry reward model. The five wins range from 0.577 to 0.732, the sixth entry (0.492) is a statistical tie, and no training radius is best for every judge: the smaller wins under DeepSeek-R1, the larger under Gemma-4 and Nemotron-3. Appendix F.8 reports additional results with a Bradley-Terry nominal kernel across both fine-tuning backbones.

Table 1: Win rate against NashMD under external LLM judges, evaluated on 1000 TL;DR prompts, with a non-BT (PairRM) nominal kernel.

Fine-Tuning Model	ρ_{train}	**Win Rate under Evaluation Model**		
		DeepSeek-R1	Gemma-4	Nemotron-3
Qwen-1.5B	0.01	**0.708**	**0.577**	0.492
	0.05	**0.625**	**0.615**	**0.732**

7 Conclusion

We studied preference alignment under uncertainty in the pairwise preference kernel. Robust Nash Alignment maximizes a KL-regularized worst-case payoff over a divergence-based ambiguity set, with separate regularization strengths for the leader and follower. We introduced a four-player proxy and established its approximation error and convergence under a joint-convexity condition. For fixed truncation, the averaged proxy gap is $O(1/T)$; with truncation of order $T^{-1/2}$, the averaged policy has robust suboptimality $O(T^{-1/2}) + \rho\gamma_{\min}$ over the prescribed leader class. The experiments illustrate practical variants under preference perturbations. These results provide a basis for analyzing alignment with uncertain pairwise preferences.

Acknowledgments and Disclosure of Funding

Shihab Ahmed, Debamita Ghosh, and Yue Wang are partially supported by an Amazon Research Award, Fall 2025, by DARPA under Agreement No. HR00112490427, and NSF ECCS-2548745, ECCS-2621991. Alvaro Velasquez recognizes support from DARPA award number HR0011261E001.

References

[1] M. Aghassi and D. Bertsimas. Robust game theory. *Mathematical programming*, 107(1):231–273, 2006.

[2] S. M. Ananthanarayanan and C. Kroer. Computing the Optimal Distributionally-Robust Strategy to Commit To. *arXiv preprint arXiv:2209.07647*, 2022.

[3] M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A General Theoretical Paradigm to Understand Learning from Human Preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

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

[5] D. Calandriello, D. Guo, R. Munos, M. Rowland, Y. Tang, B. A. Pires, P. H. Richemond, C. L. Lan, M. Valko, T. Liu, et al. Human Alignment of Large Language Models through Online Preference Optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

[6] S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. S. Bedi, and M. Wang. MaxMin-RLHF: Alignment with Diverse Human Preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[7] C.-K. Chiang, T. Yang, C.-J. Lee, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online Optimization with Gradual Variations. In *Conference on Learning Theory*, pages 6–1. JMLR Workshop and Conference Proceedings, 2012.

[8] K. Chidambaram, K. V. Seetharaman, and V. Syrgkanis. Direct Preference Optimization with Unobserved Preference Heterogeneity: The Necessity of Ternary Preferences. *arXiv preprint arXiv:2510.15716*, 2025.

[9] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep Reinforcement Learning from Human Preferences. *Advances in neural information processing systems*, 30, 2017.

[10] X. Chu, Z. Zhang, T. Jia, and Y. Jin. Stackelberg Self-Annotation: A Robust Approach to Data-Efficient LLM Alignment. *arXiv preprint arXiv:2502.18099*, 2025.

[11] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. Scaling Instruction-Finetuned Language Models, 2022.

[12] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025.

[13] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. KTO: Model Alignment as Prospect Theoretic Optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[14] H. Furuta, K.-H. Lee, S. S. Gu, Y. Matsuo, A. Faust, H. Zen, and I. Gur. Geometric-Averaged Preference Optimization for Soft Preference Labels. *Advances in Neural Information Processing Systems*, 37:57076–57114, 2024.

[15] J. Hejna, R. Rafailov, H. Sikchi, C. Finn, S. Niekum, W. B. Knox, and D. Sadigh. Contrastive Preference Learning: Learning from Human Feedback without RL. *arXiv preprint arXiv:2310.13639*, 2023.

[16] I. Hong, Z. Li, A. Bukharin, Y. Li, H. Jiang, T. Yang, and T. Zhao. Adaptive Preference Scaling for Reinforcement Learning with Human Feedback. *Advances in Neural Information Processing Systems*, 37:107249–107269, 2024.

[17] J. Hong, N. Lee, and J. Thorne. ORPO: Monolithic Preference Optimization without Reference Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024.

[18] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer Science & Business Media, 2013.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021.

[20] G. N. Iyengar. Robust Dynamic Programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[21] D. Jiang, X. Ren, and B. Y. Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada, 2023. Association for Computational Linguistics.

[22] Y. Liu, H. Xu, S.-J. S. Yang, and J. Zhang. Distributionally robust equilibrium for continuous games: Nash and Stackelberg models. *European Journal of Operational Research*, 265(2):631–643, 2018.

[23] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[24] D. Mandal, P. Sasnauskas, and G. Radanovic. Distributionally Robust Reinforcement Learning with Human Feedback. *arXiv preprint arXiv:2503.00539*, 2025.

[25] J. Mcmahan, G. Artiglio, and Q. Xie. Roping in uncertainty: Robustness and regularization in markov games. In *International Conference on Machine Learning*, pages 35267–35295. PMLR, 2024.

[26] Y. Meng, M. Xia, and D. Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward . *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[27] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[28] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[29] R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, C. Fiegel, et al. Nash Learning from Human Feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[30] A. Nayak, T. Yang, O. Yagan, G. Joshi, and Y. Chi. Achieving Logarithmic Regret in KL-Regularized Zero-Sum Markov Games. *arXiv preprint arXiv:2510.13060*, 2025.

[31] A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[32] NVIDIA, :, A. S. Deshmukh, K. Chumachenko, T. Rintamaki, M. Le, T. Poon, D. M. Taheri, I. Karmanov, G. Liu, J. Seppanen, A. Goel, M. Ranzinger, G. Heinrich, G. Chen, L. Voegtle, P. Fischer, T. Roman, K. Sapra, C. McCarthy, S. Zhang, F. Liu, H. Ye, Y. Dong, M. Liu, Y. Peng, P. Zelasko, Z. Chen, N. R. Koluguri, N. Tadevosyan, L. Grigoryan, E. H. Asl, P. Biswas, L. Tavabi, Y. Su, Z. Yu, P. Jin, A. Milesi, N. Haber, Y. Xu, S. Amiraslani, N. Mulepati, E. Tramel, J. Jung, X. Lu, B. Cui, J. Xu, Z. Li, S. Wang, Y. Kuang, S. Zhang, H. Yang, B. Li, H. Yin, S. Han, P. Molchanov, A. Renduchintala, C. Wang, D. Mosallanezhad, S. Singhal, L. Vega, K. Cheung, S. Ghosh, Y. Zhang, A. Bukharin, V. Srinivasan, J. Greco, A. Manoel, M. V. Segbroeck, S. Panguliri, R. Watve, D. Kakwani, S. Pachori, J. Glick, R. Sri-Tharan, A. Zaman, K. Nguyen, S. Chen, J. Fang, Q. Miao, W. Zhou, Y. Wang, Z. P. Bhat, V. Praveen, A. Jain, R. Arunachalam, T. Kornuta, A. Sharabiani, A. Shen, W. Huang, Y.-F. Wu, A. R. Ghias, H. Li, B. Yu, N. Tajbakhsh, C. Cui, W. Gao, L. Ding, T. Kong, M. Kilaru, A. Bhiwandiwalla, M. Wawrzos, D. Korzekwa, P. Ribalta, G. Chlebus, B. Nushi, E. Dobrowolska, M. J. Mikulski, K. Dhawan, S. Huang, J. Balam, Y. Wang, N. Karpov, V. Mendelev, G. Zelenfroynd, M. Mkrtchyan, Q. Miao, O. Almog, B. Pawar, R. Shivbhakta, S. Sabnis, A. Sharabiani, N. Habibi, G. Venkataramani, P. Peng, P. Rodney, S. Panev, R. Mazzarese, N. Liu, M. Fukuyama, A. Skliar, R. Waleffe, D. Riach, Y. Zou, J. Hu, H. Zhang, B. Xu, Y. Yang, Z. Ahmed, A. Milesi, C. del Mundo, C. Voegele, Z. Cheng, N. Assaf, A. Skliar, D. Afrimi, N. Bagrov, R. Zilberstein, O. Masad, E. Khvedchenia, N. Bagrov, B. Tymchenko, T. Asida, D. Afrimi, P. Mannan, V. Cui, M. Evans, K. Luna, J. Lou, P. Xu, G. Huang, N. Habibi, M. Boone, P. Thalasta, A. Adesoba, D. Yared, C. Parisien, L. Derczynski, S. Ghosh, W. Feely, M. Schaffer, R. Sri-Tharan, J. Glick, B. Simkin, G. Zelenfroynd, T. Grzegorzek, R. Garg, A. Jhunjhunwala, S. Kolchenko, F. Memarian, H. Kumar, S. Kumar, I. Hulseman, A. Shah, K. Briski, P. Subramanian, J. Conway, U. Karpas, J. P. Scowcroft, A. Surla, S. Ammireddy, E. Evans, J. Oliver, T. Balough, C.-C. Chen, S. Bhaskar, A. Rico, B. Sadeghi, S. Mard, K. Cheung, M. Price, L. Sleiman, S. Kaji, W. Helmholz, W. Quan, M. Lightstone, J. Cohen, J. Zhang, O. Kuchaiev, B. Ginsburg, J. Kautz, E. Long, M. Shoeybi, M. Patwary, O. Olabiyi, A. Tao, B. Catanzaro, and U. Karpas. Nemotron 3 nano omni: Efficient and open multimodal intelligence, 2026.

[33] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[34] B. Pásztor, T. K. Buening, and A. Krause. Stackelberg Learning from Human Feedback: Preference Optimization as a Sequential Game. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025.

[35] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[36] H. Rahimian and S. Mehrotra. Distributionally Robust Optimization: A Review. *arXiv preprint arXiv:1908.05659*, 2019.

[37] A. Rakhlin and K. Sridharan. Online Learning with Predictable Sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013.

[38] S. Sagawa, P. W. Koh, T. Hashimoto, and P. Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[39] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[40] A. Sharifnassab, S. Salehkaleybar, S. Ghiassian, S. Kanoria, and D. Schuurmans. Soft Preference Optimization: Aligning Language Models to Expert Distributions. *arXiv preprint arXiv:2405.00747*, 2024.

[41] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with Human Feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[42] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast Convergence of Regularized Learning in Games. *Advances in Neural Information Processing Systems*, 28, 2015.

[43] Y. Tang, Z. D. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Á. Pires, and B. Piot. Generalized Preference Optimization: A Unified Approach to Offline Alignment. *arXiv preprint arXiv:2402.05749*, 2024.

[44] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. Gemma: Open Models Based on Gemini Research and Technology, 2024.

[45] L. L. Thurstone. Psychophysical Analysis. *The American journal of psychology*, 38(3):368–389, 1927.

[46] D. Tiapkin, D. Calandriello, D. Belomestny, E. Moulines, A. Naumov, K. Rasul, M. Valko, and P. Menard. Proximal Point Nash Learning from Human Feedback. *arXiv preprint arXiv:2505.19731*, 2025.

[47] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Linear Last-iterate Convergence in Constrained Saddle-point Optimization. *arXiv preprint arXiv:2006.09517*, 2020.

[48] W. Wiesemann, D. Kuhn, and B. Rustem. Robust Markov Decision Processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[49] J. Wu, Y. Xie, Z. Yang, J. Wu, J. Chen, J. Gao, B. Ding, X. Wang, and X. He. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. *arXiv preprint arXiv:2407.07880*, 2024.

[50] Z. Xu, S. Vemuri, K. Panaganti, D. Kalathil, R. Jain, and D. Ramachandran. Robust LLM Alignment via Distributionally Robust Direct Preference Optimization. *arXiv preprint arXiv:2502.01930*, 2025.

[51] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[52] C. Ye, W. Xiong, Y. Zhang, H. Dong, N. Jiang, and T. Zhang. Online Iterative Reinforcement Learning from Human Feedback with General Preference Model. *Advances in Neural Information Processing Systems*, 37:81773–81807, 2024.

[53] K. Zhang, T. Sun, Y. Tao, S. Genc, S. Mallya, and T. Basar. Robust multi-agent reinforcement learning with model uncertainty. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[54] Y. Zhang, D. Yu, B. Peng, L. Song, Y. Tian, M. Huo, N. Jiang, H. Mi, and D. Yu. Iterative Nash Policy Optimization: Aligning LLMs with General Preferences via No-Regret Learning. In *International Conference on Learning Representations (ICLR)*, 2025. General-preference RLHF as a two-player game; OMD-based last-iterate convergence; KL-regularized formulation.

[55] Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. SLiC-HF: Sequence Likelihood Calibration with Human Feedback. *arXiv preprint arXiv:2305.10425*, 2023.

[56] R. Zhou, M. Fazel, and S. S. Du. Extragradient Preference Optimization (EGPO): Beyond Last-Iterate Convergence for Nash Learning from Human Feedback. *arXiv preprint arXiv:2503.08942*, 2025.

Appendix

Table of Contents

A Related Work . . . 16

B Main Results . . . 17

B.1 Algorithm Update Rules . . . 17

B.2 Proof of Theorem 1 . . . 17

B.3 Proof of Theorem 2 . . . 18

B.4 Proof of Theorem 3 . . . 21

B.5 Computational Cost and Dependence on the Parameters . . . 24

C Comparisons with Standard NLHF . . . 25

D Auxiliary Lemmas . . . 25

E IMDb Continuation Task: Experiment Details . . . 30

E.1 Nominal Preference and Annotator Groups . . . 30

E.2 Optimization Algorithms . . . 30

E.3 Hyperparameters . . . 31

E.4 Results . . . 31

F TL;DR Summarization Task: Experiment Details . . . 33

F.1 From Algorithm 1 to the sampled implementation . . . 33

F.2 Fine-Tuning Models . . . 33

F.3 Nominal Preference Proxy . . . 33

F.4 Adversary Proxy . . . 34

F.5 Losses . . . 34

F.6 Evaluation with LLM Judges . . . 34

F.7 Non-BT Nominal Kernel: Construction and Diagnostic . . . 35

F.8 Additional Results with a Bradley-Terry Nominal Kernel . . . 35

G LLM Judge Reasoning Examples . . . 36

A Related Work

Reward-based RLHF and direct alignment. A large body of work on alignment starts from the RLHF pipeline: collect pairwise comparisons, fit a reward model, and optimize a policy against that reward subject to a regularization term that keeps the policy close to a reference model [9, 41, 33]. This paradigm has been highly influential in large-scale language-model alignment, but it typically treats the preference signal as a nominal target and often relies, either explicitly or implicitly, on a latent-utility view of pairwise comparisons. Motivated by the computational and modeling limitations of reward-model-based RLHF, a broad line of direct or reward-free preference optimization methods has emerged, including DPO, SLiC-HF, CPL, KTO, ORPO, SimPO, SPO, and adaptive preference scaling [35, 55, 15, 13, 17, 26, 40, 16]. These methods substantially improve the practical pipeline of preference optimization, but they generally continue to optimize against a fixed nominal preference estimate rather than an ambiguity set of plausible preference kernels.

General-preference and game-theoretic alignment. Recent work has argued that pairwise human preferences should be modeled directly rather than compressed into a scalar reward function [3]. This viewpoint leads naturally to game-theoretic formulations of alignment. In particular, Nash Learning from Human Feedback (NLHF) formulates preference alignment as a regularized two-player game and introduces Nash-MD as a no-regret algorithm for the induced general-preference game [29]. Subsequent works have further developed this perspective, including online and offline algorithms under general preference models and policy-loss implementations based on no-regret learning [52, 54, 30, 56]. More recently, sequential leader–follower formulations of preference optimization have also begun to appear, including Stackelberg-style alignment methods and Stackelberg game variants for data-efficient alignment [34, 10]. Our work is closest in spirit to this line of research, but differs in one central respect: we do not assume that the learned preference game is exact. Instead, we explicitly model uncertainty in preference kernels and optimize against the worst kernel in an ambiguity set. This creates a coupled inner optimization problem, discussed in Appendix C.

Soft labels, heterogeneous preferences, and population shift. Another line of work emphasizes that preference data are often soft, noisy, and heterogeneous across annotators or user groups. Recent studies incorporate soft preference labels or pairwise confidence information into direct alignment objectives [16, 40, 14], while other work explicitly models hidden annotator types or heterogeneous preference populations [8]. These papers highlight an important limitation of nominal preference optimization: even if one optimizes the nominal objective correctly, the aggregated preference signal may still be unstable because it hides substantial variation across annotators or deployment populations. Our formulation is complementary to this line of work. Rather than only modifying the loss to account for label softness or latent subgroups, we directly treat the learned pairwise preference kernel as uncertain and seek a policy with a worst-case performance guarantee over that uncertainty set.

Robust alignment and distributionally robust preference optimization. Our formulation is also related to the broader literature on distributionally robust optimization and robust sequential decision making, where ambiguity sets are used to hedge estimation error and distribution shift [38, 20, 31]. In the alignment literature, recent work has started to robustify reward-based RLHF or DPO-style objectives against prompt-distribution shift, noisy preference data, or uncertain preference strength [16, 6, 49, 50, 24]. These methods are close in motivation to ours, but they robustify different objects: prompt distributions, training examples, or nominal pairwise losses. Our approach directly robustifies the *pairwise preference kernel itself* inside a general-preference game. This distinction is important because kernel uncertainty directly captures ambiguity in preferences, which is the primitive object in general-preference alignment.

Robust game theory and robust Markov games. A separate line of work studies robustness to payoff or model uncertainty directly in game-theoretic solution concepts, rather than in a single-agent decision problem. This includes robust game theory under set-based payoff uncertainty [1], distributionally robust Nash and Stackelberg equilibrium models under exogenous distributional uncertainty [22], and distributionally robust commitment to a strategy when the follower's utility is known only up to a finite set of candidate models [2]. Robust Markov games extend this to multi-stage settings with uncertain rewards or transitions [53, 25]. Our setting differs from this literature in two respects. First, robust Markov games study equilibria under model uncertainty, whereas our objective evaluates one learned policy against a competing policy and a jointly chosen adverse kernel. Second, a single Bernoulli-KL budget couples response pairs within each prompt; it

does not provide the componentwise decomposition used by rectangular robust dynamic programs. Existing robust-game and robust-Markov-game results therefore do not directly provide our proxy formulation or its convergence guarantee.

B Main results

B.1 Algorithm Update Rules

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\epsilon} \left\{ \langle \widehat{g}_t^{\pi}, \pi \rangle - \frac{1}{\eta} D_{\mathrm{KL}}(\pi \| \pi_t) \right\}, \tag{26}$$

$$\gamma_{t+1} = \mathrm{Proj}_{\Gamma}\big(\gamma_t + \eta \widehat{g}_t^{\gamma}\big), \tag{27}$$

$$\nu_{t+1} = \arg\min_{\nu \in \mathcal{N}_\beta} \left\{ \langle \widehat{g}_t^{\nu}, \nu \rangle + \frac{1}{\eta} D_{\mathrm{KL}}(\nu \| \nu_t) \right\}, \tag{28}$$

$$p_{t+1} = \arg\min_{p \in P_\delta} \left\{ \langle \widehat{g}_t^{p}, p \rangle + \frac{1}{\eta} D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p_t) \right\}. \tag{29}$$

B.2 Proof of Theorem 1

Proof. Fix $\pi \in \Pi_\epsilon$. Since the feasible set in (15) is a subset of the feasible set in (13), the lower bound

$$g_H^{\delta}(\pi; \rho) \geq g_H(\pi; \rho)$$

is immediate. It remains to show the upper bound. Fix $\nu \in \Delta_m$, and let

$$K(\rho) := \left\{ p \in [0,1]^d : \; D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) \leq \rho \right\}.$$

Take any $p \in K(\rho)$, and clip it coordinatewise into P_δ: $\widetilde{p} := \mathrm{Proj}_{P_\delta}(p)$. Because $p^\star \in P_\delta$, each scalar $\widetilde{p}_e$ lies on the line segment joining p_e and $p_e^\star$. For each coordinate e, the scalar function

$$u \mapsto u \log \frac{u}{p_e^\star} + (1-u) \log \frac{1-u}{1-p_e^\star}$$

is convex and has its unique minimum at $u = p_e^\star$. Hence clipping toward $p_e^\star$ cannot increase the coordinatewise Bernoulli-KL divergence, so

$$D_{\mathrm{KL}}^{\mathrm{Bern}}(\widetilde{p} \| p^\star) \leq D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) \leq \rho.$$

Thus $\widetilde{p}$ is feasible for the truncated problem.

Now $J(\pi, \nu; p)$ depends on p only through the linear term $\sum_{(i,j)\in E} w_{ij}(\pi,\nu)\, p_{ij}$, where $w_{ij}(\pi,\nu) := \pi_i \nu_j - \pi_j \nu_i$; indeed, by antisymmetry, $\sum_{i,j} \pi_i \nu_j P_p(i \succ j) = \frac{1}{2}\langle \pi, \nu \rangle + \sum_{i<j} \pi_j \nu_i + \sum_{(i,j)\in E} w_{ij}(\pi,\nu)\, p_{ij}$. Therefore

$$|J(\pi,\nu;\widetilde{p}) - J(\pi,\nu;p)| = \left| \sum_{(i,j)\in E} w_{ij}(\pi,\nu)\, (\widetilde{p}_{ij} - p_{ij}) \right| \leq \|\widetilde{p} - p\|_\infty \sum_{(i,j)\in E} |w_{ij}(\pi,\nu)|.$$

Since each coordinate is clipped by at most δ, we have

$$\|\widetilde{p} - p\|_\infty \leq \delta.$$

Also,

$$\sum_{(i,j)\in E} |w_{ij}(\pi,\nu)| \leq \sum_{(i,j)\in E} (\pi_i \nu_j + \pi_j \nu_i) = \sum_{i \neq j} \pi_i \nu_j = 1 - \sum_{i=1}^{m} \pi_i \nu_i \leq 1.$$

Hence

$$J(\pi,\nu;\widetilde{p}) \leq J(\pi,\nu;p) + \delta.$$

Taking the minimum first over $p \in K(\rho)$, then over $\nu \in \Delta_m$, yields

$$g_H^{\delta}(\pi;\rho) \leq g_H(\pi;\rho) + \delta.$$

This proves the claim $0 \le g_H^\delta(\pi;\rho) - g_H(\pi;\rho) \le \delta$. Maximizing over $\pi \in \Pi_\epsilon$ gives $0 \le V^\delta(\rho) - V(\rho) \le \delta$.

Using the above two claims, we have

$$V(\rho) \le V^\delta(\rho) = g_H^\delta(\pi_\delta^\star;\rho) \le g_H(\pi_\delta^\star;\rho) + \delta.$$

Rearranging gives the claim. □

B.3 Proof of Theorem 2

Proof. Fix $\pi \in \Pi_\epsilon$, and define

$$\phi_\pi(\nu,\gamma) := \min_{p\in P_\delta} F(\pi,\gamma,\nu,p).$$

Step 1: reduction to $\mathcal{N}_\beta$ and nested strong duality. For fixed (π,p), define

$$b_j(\pi,p) := \sum_{i=1}^m \pi(i)\, P_p(i \succ j), \qquad j = 1,\dots,m.$$

Since $P_p(i \succ j) \in [0,1]$ and $\sum_i \pi(i) = 1$, we have $b_j(\pi,p) \in [0,1]$ for all j.

For fixed (π,p), the follower subproblem is

$$\min_{\nu\in\Delta_m} \left\{ \sum_{j=1}^m \nu(j)\, b_j(\pi,p) + \tau_\nu D_{\mathrm{KL}}(\nu\|\pi_{\mathrm{ref}}) \right\}.$$

Its unique minimizer is the Gibbs distribution

$$\nu_j^{\mathrm{br}}(\pi,p) = \frac{\pi_{\mathrm{ref}}(j) e^{-b_j(\pi,p)/\tau_\nu}}{\sum_{\ell=1}^m \pi_{\mathrm{ref}}(\ell) e^{-b_\ell(\pi,p)/\tau_\nu}}.$$

Because $b_j(\pi,p) \in [0,1]$, we have $e^{-1/\tau_\nu} \le e^{-b_j(\pi,p)/\tau_\nu} \le 1$. Let

$$Z(\pi,p) := \sum_{\ell=1}^m \pi_{\mathrm{ref}}(\ell) e^{-b_\ell(\pi,p)/\tau_\nu}.$$

Since $\sum_\ell \pi_{\mathrm{ref}}(\ell) = 1$ and each exponential factor is at most 1, we have $Z(\pi,p) \le 1$. Therefore

$$\nu_j^{\mathrm{br}}(\pi,p) = \pi_{\mathrm{ref}}(j) \frac{e^{-b_j(\pi,p)/\tau_\nu}}{Z(\pi,p)} \ge \pi_{\mathrm{ref}}(j) e^{-1/\tau_\nu}.$$

Under Assumption 2-(2), $\beta \le e^{-1/\tau_\nu}$, this implies

$$\nu^{\mathrm{br}}(\pi,p) \in \mathcal{N}_\beta.$$

Hence, for every (π,p),

$$\min_{\nu\in\Delta_m} J(\pi,\nu;p) = \min_{\nu\in\mathcal{N}_\beta} J(\pi,\nu;p). \tag{30}$$

The same identity holds with $J(\pi,\nu;p)$ replaced by $F(\pi,\gamma,\nu,p)$, since the dual term $\gamma(D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star) - \rho)$ does not depend on ν.

Now fix $\nu \in \mathcal{N}_\beta$. The p-subproblem

$$\min_{\substack{p\in P_\delta \\ D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star)\le\rho}} J(\pi,\nu;p)$$

is a convex program: the objective is affine in p, the feasible set is convex, and Slater's condition holds because Assumption 2-(1) gives $p^\star \in (\delta,1-\delta)^d$ and $D_{\mathrm{KL}}^{\mathrm{Bern}}(p^\star\|p^\star) = 0 < \rho$. Therefore strong duality yields

$$\min_{\substack{p\in P_\delta \\ D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star)\le\rho}} J(\pi,\nu;p) = \max_{\gamma\ge 0} \min_{p\in P_\delta} F(\pi,\gamma,\nu,p).$$

By Lemma 1, every scalar dual maximizer lies in $[0, 1/\rho]$. Assumption 2-(3) therefore permits restriction to $[0, \gamma_{\max}]$. Consequently,

$$\min_{\substack{p \in P_\delta \\ D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star) \le \rho}} J(\pi, \nu; p) = \max_{\gamma \in [0, \gamma_{\max}]} \min_{p \in P_\delta} F(\pi, \gamma, \nu, p).$$

Combining this with (30), we obtain

$$g_H^\delta(\pi; \rho) = \min_{\nu \in \mathcal{N}_\beta} \max_{\gamma \in [0, \gamma_{\max}]} \min_{p \in P_\delta} F(\pi, \gamma, \nu, p). \tag{31}$$

Step 2: proof of the lower bound (20)**.** By definition,

$$g_{\mathrm{proxy}}^\delta(\pi; \rho) = \max_{\gamma \in \Gamma} \min_{\nu \in \mathcal{N}_\beta,\ p \in P_\delta} F(\pi, \gamma, \nu, p) = \max_{\gamma \in \Gamma} \min_{\nu \in \mathcal{N}_\beta} \phi_\pi(\nu, \gamma).$$

Since $\Gamma = [\gamma_{\min}, \gamma_{\max}] \subseteq [0, \gamma_{\max}]$,

$$g_{\mathrm{proxy}}^\delta(\pi; \rho) \le \max_{\gamma \in [0, \gamma_{\max}]} \min_{\nu \in \mathcal{N}_\beta} \phi_\pi(\nu, \gamma).$$

By weak minimax,

$$\max_{\gamma \in [0, \gamma_{\max}]} \min_{\nu \in \mathcal{N}_\beta} \phi_\pi(\nu, \gamma) \le \min_{\nu \in \mathcal{N}_\beta} \max_{\gamma \in [0, \gamma_{\max}]} \phi_\pi(\nu, \gamma).$$

Using (31), the right-hand side equals $g_H^\delta(\pi; \rho)$. Therefore

$$g_{\mathrm{proxy}}^\delta(\pi; \rho) \le g_H^\delta(\pi; \rho),$$

which proves (20).

Step 3: excluding the interval $[0, \gamma_{\min})$ **costs at most** $\rho\gamma_{\min}$**.** Fix $\nu \in \mathcal{N}_\beta$, and define

$$q_{\pi,\nu}(\gamma) := \phi_\pi(\nu, \gamma).$$

We claim that for any $0 \le a < b \le \gamma_{\max}$,

$$q_{\pi,\nu}(a) \le q_{\pi,\nu}(b) + \rho(b - a). \tag{32}$$

Indeed, let $p_b \in \arg\min_{p \in P_\delta} F(\pi, b, \nu, p)$. Then

$$\begin{aligned} q_{\pi,\nu}(a) &\le F(\pi, a, \nu, p_b) \\ &= F(\pi, b, \nu, p_b) + (a - b)\big(D_{\mathrm{KL}}^{\mathrm{Bern}}(p_b\|p^\star) - \rho\big) \\ &= q_{\pi,\nu}(b) + (b - a)\big(\rho - D_{\mathrm{KL}}^{\mathrm{Bern}}(p_b\|p^\star)\big) \\ &\le q_{\pi,\nu}(b) + \rho(b - a), \end{aligned}$$

since $D_{\mathrm{KL}}^{\mathrm{Bern}}(p_b\|p^\star) \ge 0$. This proves (32).

Now let $\gamma^\sharp \in [0, \gamma_{\max}]$ be a maximizer of $q_{\pi,\nu}$ over $[0, \gamma_{\max}]$. If $\gamma^\sharp \in \Gamma$, then

$$\max_{\gamma \in [0, \gamma_{\max}]} q_{\pi,\nu}(\gamma) = \max_{\gamma \in \Gamma} q_{\pi,\nu}(\gamma).$$

If $\gamma^\sharp < \gamma_{\min}$, then applying (32) with $a = \gamma^\sharp$ and $b = \gamma_{\min}$ yields

$$q_{\pi,\nu}(\gamma^\sharp) \le q_{\pi,\nu}(\gamma_{\min}) + \rho(\gamma_{\min} - \gamma^\sharp) \le \max_{\gamma \in \Gamma} q_{\pi,\nu}(\gamma) + \rho\gamma_{\min}.$$

Hence, in all cases,

$$\max_{\gamma \in [0, \gamma_{\max}]} q_{\pi,\nu}(\gamma) \le \max_{\gamma \in \Gamma} q_{\pi,\nu}(\gamma) + \rho\gamma_{\min}.$$

Taking the minimum over $\nu \in \mathcal{N}_\beta$, we obtain

$$g_H^\delta(\pi; \rho) \le \min_{\nu \in \mathcal{N}_\beta} \max_{\gamma \in \Gamma} \phi_\pi(\nu, \gamma) + \rho\gamma_{\min}. \tag{33}$$

Step 4: under Assumption 2-(4), the minimax defect on Γ vanishes. Assume now Assumption 2-(4), i.e.

$$\tau_\nu \gamma_{\min} \geq \frac{1}{4}.$$

We show that $\phi_\pi(\cdot, \cdot)$ is convex in ν and concave in γ on $\mathcal{N}_\beta \times \Gamma$.

Concavity in γ. For fixed ν, the function

$$\gamma \mapsto F(\pi, \gamma, \nu, p)$$

is affine for every fixed p, so $\phi_\pi(\nu, \cdot)$, being the pointwise infimum of affine functions, is concave on Γ.

Convexity in ν. Fix $\gamma \in \Gamma$. We claim that $(\nu, p) \mapsto F(\pi, \gamma, \nu, p)$ is jointly convex on $\mathcal{N}_\beta \times P_\delta$. Write

$$W(\pi, \nu, p) = \sum_{j=1}^{m} \nu(j)\, b_j(\pi, p), \qquad b_j(\pi, p) = \sum_{i=1}^{m} \pi(i) P_p(i \succ j).$$

Let (ν, p) and (ν', p') be two points in $\mathcal{N}_\beta \times P_\delta$. Since W is bilinear in (ν, p),

$$W(\pi, \nu', p') - W(\pi, \nu, p) - \langle \nabla_\nu W(\pi, \nu, p), \nu' - \nu \rangle - \langle \nabla_p W(\pi, \nu, p), p' - p \rangle = \langle \nu' - \nu,\ b(\pi, p') - b(\pi, p) \rangle.$$

For each coordinate j,

$$|b_j(\pi, p') - b_j(\pi, p)| \leq \sum_{i=1}^{m} \pi(i)\, |P_{p'}(i \succ j) - P_p(i \succ j)| \leq \|p' - p\|_2,$$

hence

$$\left| \langle \nu' - \nu,\ b(\pi, p') - b(\pi, p) \rangle \right| \leq \|\nu' - \nu\|_1 \, \|p' - p\|_2.$$

Therefore,

$$\begin{aligned} F(\pi, \gamma, \nu', p') \geq\ & F(\pi, \gamma, \nu, p) + \langle \nabla_\nu F(\pi, \gamma, \nu, p), \nu' - \nu \rangle + \langle \nabla_p F(\pi, \gamma, \nu, p), p' - p \rangle \\ & + \tau_\nu D_{\mathrm{KL}}(\nu' \| \nu) + \gamma D_{\mathrm{KL}}^{\mathrm{Bern}}(p' \| p) - \|\nu' - \nu\|_1 \, \|p' - p\|_2. \end{aligned}$$

Using Pinsker's inequalities

$$D_{\mathrm{KL}}(\nu' \| \nu) \geq \frac{1}{2} \|\nu' - \nu\|_1^2, \qquad D_{\mathrm{KL}}^{\mathrm{Bern}}(p' \| p) \geq 2 \|p' - p\|_2^2,$$

we obtain

$$\tau_\nu D_{\mathrm{KL}}(\nu' \| \nu) + \gamma D_{\mathrm{KL}}^{\mathrm{Bern}}(p' \| p) - \|\nu' - \nu\|_1 \, \|p' - p\|_2 \geq \frac{\tau_\nu}{2} s^2 + 2\gamma t^2 - st,$$

where $s = \|\nu' - \nu\|_1$ and $t = \|p' - p\|_2$. Since $\gamma \geq \gamma_{\min}$ and $\tau_\nu \gamma_{\min} \geq 1/4$,

$$\frac{\tau_\nu}{2} s^2 + 2\gamma t^2 - st = \left(\sqrt{\frac{\tau_\nu}{2}}\, s - \sqrt{2\gamma}\, t \right)^2 + 2 \left(\sqrt{\tau_\nu \gamma} - \frac{1}{2} \right) st \geq 0.$$

Thus $(\nu, p) \mapsto F(\pi, \gamma, \nu, p)$ is jointly convex. Partial minimization over the convex set P_δ preserves convexity: for minimizers p_1, p_2 at ν_1, ν_2, evaluate at their convex combination and apply joint convexity. Hence $\phi_\pi(\cdot, \gamma)$ is convex. Continuity of F on the compact domain makes ϕ_π continuous. Since $\mathcal{N}_\beta$ and Γ are compact and convex, Sion's minimax theorem applies:

$$\min_{\nu \in \mathcal{N}_\beta} \max_{\gamma \in \Gamma} \phi_\pi(\nu, \gamma) = \max_{\gamma \in \Gamma} \min_{\nu \in \mathcal{N}_\beta} \phi_\pi(\nu, \gamma) = g_{\mathrm{proxy}}^\delta(\pi; \rho).$$

Substituting this into (33) yields

$$g_H^\delta(\pi; \rho) \leq g_{\mathrm{proxy}}^\delta(\pi; \rho) + \rho \gamma_{\min}.$$

Combining with (20), we conclude

$$0 \leq g_H^\delta(\pi; \rho) - g_{\mathrm{proxy}}^\delta(\pi; \rho) \leq \rho \gamma_{\min},$$

which proves the first inequality in (21).

Step 5: pass from pointwise values to optimal values. By definition,

$$V^\delta(\rho) = \max_{\pi \in \Pi_\epsilon} g_H^\delta(\pi;\rho), \qquad V_{\mathrm{proxy}}^\delta(\rho) = \max_{\pi \in \Pi_\epsilon} g_{\mathrm{proxy}}^\delta(\pi;\rho).$$

For any two functions f, g on the same domain,

$$\max_\pi f(\pi) - \max_\pi g(\pi) \le \max_\pi (f(\pi) - g(\pi)).$$

Applying this with $f(\pi) = g_H^\delta(\pi;\rho)$ and $g(\pi) = g_{\mathrm{proxy}}^\delta(\pi;\rho)$, and using the pointwise bound just proved, we obtain

$$V^\delta(\rho) - V_{\mathrm{proxy}}^\delta(\rho) \le \max_{\pi \in \Pi_\epsilon} \big(g_H^\delta(\pi;\rho) - g_{\mathrm{proxy}}^\delta(\pi;\rho)\big) \le \rho\gamma_{\min}.$$

This proves the second inequality in (21).

□

B.4 Proof of Theorem 3

For completeness, we state Theorem 3 with the explicit smoothness constant.

Theorem 4 (Convergence and robust guarantee with explicit truncation dependence)**.** *Under Assumption 2, let* $\pi_{\mathrm{ref,min}} := \min_i \pi_{\mathrm{ref},i}$ *and*

$$L_\delta := \max\left\{1, \frac{\tau_\pi}{\epsilon\pi_{\mathrm{ref,min}}}, \frac{\tau_\nu}{\beta\pi_{\mathrm{ref,min}}}, \frac{\gamma_{\max}}{\delta(1-\delta)}, \sqrt{d}, 2\sqrt{d}\log\frac{1-\delta}{\delta}\right\}.$$

Run Algorithm 1 with $\eta = 1/(8L_\delta)$*. Use the diameters,* H_δ*,* R*, and* C_δ *defined in Theorem 3, and put* $B_{T,\delta} := C_\delta/T$*. Then*

$$\mathrm{Gap}(\bar z_T) \le B_{T,\delta}, \tag{34}$$

$$V(\rho) - g_H(\bar\pi_T;\rho) \le B_{T,\delta} + \rho\gamma_{\min} + \delta. \tag{35}$$

For the fixed initialization and truncation schedule in Theorem 3, $C_\delta = O(1/\delta)$ *and the robust suboptimality is* $O(T^{-1/2}) + \rho\gamma_{\min}$*.*

Proof. We first bound the proxy saddle gap, transfer this bound to the hard objective, and then track the constants as the truncation decreases.

Step 1: proxy-game convergence with explicit L_δ. Let

$$x_t := (\pi_t, \gamma_t) \in X, \qquad y_t := (\nu_t, p_t) \in Y, \qquad z_t = (x_t, y_t),$$

and define the averaged iterates

$$\bar x_T := \frac{1}{T}\sum_{t=1}^T x_t, \qquad \bar y_T := \frac{1}{T}\sum_{t=1}^T y_t.$$

For any $x \in X$, define

$$A_T(x) := \sum_{t=1}^T \big(F(x, y_t) - F(x_t, y_t)\big),$$

and for any $y \in Y$, define

$$B_T(y) := \sum_{t=1}^T \big(F(x_t, y_t) - F(x_t, y)\big).$$

Step 4 of the proof of Theorem 2 establishes convexity in $y = (\nu, p)$. For fixed y, F is jointly concave in $x = (\pi, \gamma)$ because its only nonlinear term in x is $-\tau_\pi D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}})$. Applying these first-order inequalities and Lemma 2 gives

$$A_T(x) + B_T(y) \le \frac{D_X(x, x_1) + D_Y(y, y_1)}{\eta} + E_T(x, y) + V_T - \Xi_T, \tag{36}$$

where

$$D_X(x, x_1) := D_{\mathrm{KL}}(\pi\|\pi_1) + \frac{1}{2}(\gamma - \gamma_1)^2, \qquad D_Y(y, y_1) := D_{\mathrm{KL}}(\nu\|\nu_1) + D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p_1),$$

Write $D_h((x, y), (x', y')) := D_X(x, x') + D_Y(y, y')$. The endpoint contribution is

$$E_T(x, y) := -\frac{D_h((x, y), z_{T+1})}{\eta} + \langle G(z_1), z_1 - (x, y)\rangle - \langle G(z_T), z_{T+1} - (x, y)\rangle.$$

Here $G(z_0)$ in the blockwise identities denotes the initialized value $G(z_1)$, rather than an additional iterate. The variation and stability sums are

$$V_T := \eta \sum_{t=1}^{T} \Big(\|g_t^\pi - g_{t-1}^\pi\|_\infty^2 + |g_t^\gamma - g_{t-1}^\gamma|^2 + \|g_t^\nu - g_{t-1}^\nu\|_\infty^2 + \|g_t^p - g_{t-1}^p\|_2^2 \Big),$$

$$\Xi_T := \frac{1}{2\eta} \sum_{t=1}^{T} \Big(D_{\mathrm{KL}}(\pi_{t+1}\|\pi_t) + \tfrac{1}{2}|\gamma_{t+1} - \gamma_t|^2 + D_{\mathrm{KL}}(\nu_{t+1}\|\nu_t) + D_{\mathrm{KL}}^{\mathrm{Bern}}(p_{t+1}\|p_t) \Big).$$

Lemma 3 bounds the gradient variation, and Lemma 4 gives $V_T \le \Xi_T$ for $\eta = 1/(8L_\delta)$. It remains to bound E_T. Pinsker's inequalities and the Euclidean block give $D_h(u, v) \ge \frac{1}{2}\|u - v\|^2$. For $u = (x, y)$ and $a := \|u - z_{T+1}\|$, duality of the norms yields

$$E_T(x, y) \le H_\delta R - \frac{a^2}{2\eta} + H_\delta a \le H_\delta R + \frac{\eta H_\delta^2}{2}.$$

The last inequality maximizes the concave quadratic in a. Therefore

$$A_T(x) + B_T(y) \le \frac{D_X(x, x_1) + D_Y(y, y_1)}{\eta} + H_\delta R + \frac{\eta H_\delta^2}{2} \le C_\delta, \tag{37}$$

uniformly over $x \in X$ and $y \in Y$.

Now choose

$$\hat{x} \in \arg\max_{x \in X} F(x, \bar{y}_T), \qquad \hat{y} \in \arg\min_{y \in Y} F(\bar{x}_T, y).$$

Then

$$\mathrm{Gap}(\bar{z}_T) = F(\hat{x}, \bar{y}_T) - F(\bar{x}_T, \hat{y}).$$

Because $F(\cdot, y)$ is concave and $F(x, \cdot)$ is convex, Jensen's inequality gives

$$F(\hat{x}, \bar{y}_T) \le \frac{1}{T} \sum_{t=1}^{T} F(\hat{x}, y_t), \qquad F(\bar{x}_T, \hat{y}) \ge \frac{1}{T} \sum_{t=1}^{T} F(x_t, \hat{y}).$$

Hence

$$\begin{aligned} T\ \mathrm{Gap}(\bar{z}_T) &\le \sum_{t=1}^{T} \big(F(\hat{x}, y_t) - F(x_t, \hat{y}) \big) \\ &= \sum_{t=1}^{T} \big(F(\hat{x}, y_t) - F(x_t, y_t) \big) + \sum_{t=1}^{T} \big(F(x_t, y_t) - F(x_t, \hat{y}) \big) \\ &= A_T(\hat{x}) + B_T(\hat{y}). \end{aligned}$$

Applying (37) to $(\hat{x}, \hat{y})$ gives $T\,\mathrm{Gap}(\bar{z}_T) \le C_\delta$, proving (34). All maxima and minima above are attained by continuity on compact sets.

Step 2: proxy-policy guarantee. By definition of the duality gap,

$$\mathrm{Gap}(\bar{z}_T) = \max_{x' \in X} F(x', \bar{y}_T) - \min_{y' \in Y} F(\bar{x}_T, y').$$

Since $\max_{x'} F(x', \bar{y}_T) \ge F(\bar{x}_T, \bar{y}_T) = F(\bar{z}_T)$, we obtain

$$\mathrm{Gap}(\bar{z}_T) \ge F(\bar{z}_T) - \min_{y' \in Y} F(\bar{x}_T, y').$$

Rearranging,

$$\min_{y' \in Y} F(\bar{x}_T, y') \geq F(\bar{z}_T) - \mathrm{Gap}(\bar{z}_T). \tag{38}$$

Now

$$g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) = \max_{\gamma \in \Gamma} \min_{y \in Y} F((\bar{\pi}_T, \gamma), y).$$

Since $\bar{x}_T = (\bar{\pi}_T, \bar{\gamma}_T) \in X$,

$$g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) \geq \min_{y \in Y} F(\bar{x}_T, y).$$

Combining this with (38) and (34) gives

$$g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) \geq F(\bar{z}_T) - \mathrm{Gap}(\bar{z}_T) \geq F(\bar{z}_T) - B_{T,\delta}. \tag{39}$$

Step 3: transfer from the proxy game to the original hard objective. By Theorem 2, under Assumption 2 we have the pointwise bounds

$$0 \leq g^{\delta}_H(\pi; \rho) - g^{\delta}_{\mathrm{proxy}}(\pi; \rho) \leq \rho\gamma_{\min}, \qquad 0 \leq V^{\delta}(\rho) - V^{\delta}_{\mathrm{proxy}}(\rho) \leq \rho\gamma_{\min}, \qquad \forall \pi \in \Pi_{\epsilon}.$$

By Theorem 1, the truncation error satisfies

$$0 \leq g^{\delta}_H(\pi; \rho) - g_H(\pi; \rho) \leq \delta, \qquad 0 \leq V^{\delta}(\rho) - V(\rho) \leq \delta, \qquad \forall \pi \in \Pi_{\epsilon}.$$

From (39),

$$g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) \geq F(\bar{z}_T) - B_{T,\delta}.$$

Hence

$$g_H(\bar{\pi}_T; \rho) \geq g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) - \delta \geq F(\bar{z}_T) - B_{T,\delta} - \delta.$$

This gives the explicit robust-value certificate.

Next, using $V(\rho) \leq V^{\delta}(\rho)$ and $g_H(\bar{\pi}_T; \rho) \geq g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) - \delta$, we get

$$\begin{aligned} V(\rho) - g_H(\bar{\pi}_T; \rho) &\leq V^{\delta}(\rho) - g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) + \delta \\ &= \big(V^{\delta}(\rho) - V^{\delta}_{\mathrm{proxy}}(\rho)\big) + \big(V^{\delta}_{\mathrm{proxy}}(\rho) - g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho)\big) + \delta. \end{aligned}$$

By Theorem 2,

$$V^{\delta}(\rho) - V^{\delta}_{\mathrm{proxy}}(\rho) \leq \rho\gamma_{\min}.$$

Also,

$$V^{\delta}_{\mathrm{proxy}}(\rho) = \max_{\pi \in \Pi_{\epsilon}} g^{\delta}_{\mathrm{proxy}}(\pi; \rho) = \max_{x \in X} \min_{y \in Y} F(x, y),$$

so

$$V^{\delta}_{\mathrm{proxy}}(\rho) \leq \max_{x \in X} F(x, \bar{y}_T).$$

And from the definition of $g^{\delta}_{\mathrm{proxy}}$,

$$g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) \geq \min_{y \in Y} F(\bar{x}_T, y).$$

Subtracting the last two displays,

$$V^{\delta}_{\mathrm{proxy}}(\rho) - g^{\delta}_{\mathrm{proxy}}(\bar{\pi}_T; \rho) \leq \max_{x \in X} F(x, \bar{y}_T) - \min_{y \in Y} F(\bar{x}_T, y) = \mathrm{Gap}(\bar{z}_T) \leq B_{T,\delta}.$$

Combining the last three inequalities proves

$$V(\rho) - g_H(\bar{\pi}_T; \rho) \leq B_{T,\delta} + \rho\gamma_{\min} + \delta.$$

This proves (35).

Step 4: uniform constants as δ decreases. Fix the initialization in Theorem 3, and define

$$K_{\star} := \sum_{e=1}^{d} \log \frac{1}{\min\{p^{\star}_e, 1 - p^{\star}_e\}}.$$

For any $p \in [0,1]^d$, $D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p^\star) \le K_\star$: each scalar divergence is convex in its first argument, so its maximum on $[0,1]$ is at an endpoint. Consequently,

$$D_X^{\mathrm{diam}} + D_Y^{\mathrm{diam}} \le 2\log\frac{1}{\pi_{\mathrm{ref,min}}} + \frac{1}{2}\max\{\gamma_{\max}-\gamma_1, \gamma_1-\gamma_{\min}\}^2 + K_\star,$$
$$R^2 \le 8 + (\gamma_{\max}-\gamma_{\min})^2 + d.$$

These bounds are independent of δ. For the gradient bound, put

$$A_\pi := 1 + \tau_\pi(1 + \max\{\log(1/\epsilon), \log(1/\pi_{\mathrm{ref,min}})\}),$$
$$A_\nu := 1 + \tau_\nu(1 + \max\{\log(1/\beta), \log(1/\pi_{\mathrm{ref,min}})\}).$$

Lemma 6 then gives

$$H_\delta^2 \le A_\pi^2 + A_\nu^2 + (K_\star+\rho)^2 + \left(1 + 2\gamma_{\max}\sqrt{d}\log\frac{1-\delta}{\delta}\right)^2.$$

For the kernel block we used $\|w(\pi,\nu)\|_2 \le \|w(\pi,\nu)\|_1 \le 1$, proved in Appendix B.2. Thus $H_\delta = O(1+\log(1/\delta))$, whereas $L_\delta = \Theta(1/\delta)$ with all other parameters fixed. Substituting into the definition of C_δ gives $C_\delta = O(1/\delta)$. Taking $\delta_T = \min\{cT^{-1/2}, \delta_0\}$ in (35) proves the rate stated in Theorem 3. □

B.5 Computational Cost and Dependence on the Parameters

This subsection makes the cost paragraph of Section 5 explicit.

One iteration. The four gradients in Lemma 6 require two matrix–vector products with P_p, the $d = \binom{m}{2}$ weights $w_{ij} = \pi_i\nu_j - \pi_j\nu_i$, and a sum of scalar divergences. Their cost is $O(m^2)$ time and memory. The policy steps are multiplicative updates followed by KL projections onto floored simplices; a scalar bisection implements each projection. The kernel step is coordinatewise and the multiplier is clipped. Projection accuracy adds a bisection-dependent cost; at fixed numerical precision the per-iteration cost is $O(m^2)$.

Number of iterations. For exact updates, Theorem 3 gives $T = \lceil C_\delta/\varepsilon_{\mathrm{acc}}\rceil$ iterations for proxy gap at most $\varepsilon_{\mathrm{acc}}$. The constant includes the initialization and terminal-gradient terms as well as the Bregman diameters. Its smoothness component is

$$L_\delta = \max\left\{1,\ \frac{\tau_\pi}{\epsilon\,\pi_{\mathrm{ref,min}}},\ \frac{\tau_\nu}{\beta\,\pi_{\mathrm{ref,min}}},\ \frac{\gamma_{\max}}{\delta(1-\delta)},\ \sqrt{d},\ 2\sqrt{d}\log\frac{1-\delta}{\delta}\right\}.$$

For $\pi_1 = \nu_1 = \pi_{\mathrm{ref}}$ and $p_1 = p^\star \in P_\delta$ the diameters are bounded by

$$D_X^{\mathrm{diam}} \le \log\frac{1}{\pi_{\mathrm{ref,min}}} + \frac{1}{2}\max\{\gamma_{\max}-\gamma_1,\ \gamma_1-\gamma_{\min}\}^2, \qquad D_Y^{\mathrm{diam}} \le \log\frac{1}{\pi_{\mathrm{ref,min}}} + K_\star,$$

where $K_\star = \sum_e \log(1/\min\{p_e^\star, 1-p_e^\star\})$. These bounds follow from $D_{\mathrm{KL}}(\pi\|\pi_{\mathrm{ref}}) \le \log(1/\pi_{\mathrm{ref,min}})$ and the scalar divergence endpoint bound used in Appendix B.4.

Dependence on the parameters. *Truncation δ.* With the fixed initialization above, $C_\delta = O(1/\delta)$ and the truncation error is δ. Balancing C_δ/T with this error gives $\delta \asymp T^{-1/2}$, capped below the nominal kernel's distance to the boundary.

Radius ρ. Lemma 1 justifies $\gamma_{\max} \ge 1/\rho$. If $\gamma_1 \le \gamma_{\max}/2$, then $L_\delta D_X^{\mathrm{diam}} \ge \gamma_{\max}^3/(8\delta(1-\delta)) \ge 1/(8\rho^3\delta(1-\delta))$. Thus this component of the sufficient iteration bound grows at least as ρ^{-3} when $\rho \downarrow 0$ at fixed δ. This describes the bound, rather than an iteration lower bound for the algorithm.

Curvature floor $\gamma_{\min}$. The robust error bound contains $\rho\gamma_{\min}$, and the curvature condition requires $\rho\gamma_{\min} \ge \rho/(4\tau_\nu)$. This is a lower bound on the approximation allowance in our guarantee, not a lower bound on the actual policy error.

KL strengths. The leader coefficient contributes $\tau_\pi/(\epsilon\pi_{\mathrm{ref,min}})$ to L_δ. The follower floor satisfies $\beta \le e^{-1/\tau_\nu}$, so $L_\delta \ge \tau_\nu/(\beta\pi_{\mathrm{ref,min}}) \ge \tau_\nu e^{1/\tau_\nu}/\pi_{\mathrm{ref,min}}$, which grows exponentially as $\tau_\nu \downarrow 0$.

C Comparisons with Standard NLHF

Compared with standard NLHF [29] and subsequent work [54, 46, 30, 56, 52], the convergence analysis in our setting is more involved. In nominal NLHF, the preference kernel is fixed, and the optimization problem is a regularized two-player zero-sum game between the main policy and the adversarial policy. Once the game is written in saddle-point form, the KL regularizers provide strong concavity in the leader policy and strong convexity in the follower policy, so one can analyze mirror descent or optimistic no-regret dynamics directly on a fixed convex–concave game. By contrast, our robust formulation introduces uncertainty in the preference kernel itself. The resulting objective is no longer a standard two-player nominal game, but a distributionally robust max-min problem in which the adversary chooses both a competing policy and a plausible preference kernel. This creates two new difficulties that do not appear in standard NLHF.

First, the original hard-constrained problem does not retain the clean convex-concave structure of nominal NLHF. If one eliminates the follower using the Gibbs best response, the reduced objective becomes a concave minimization problem over the kernel ambiguity set, which is generally nonconvex. If one instead keeps both the competing policy and kernel, their bilinear interaction need not be jointly convex. Therefore, unlike in standard NLHF, one cannot analyze the original hard-constrained robust objective directly using a standard minimax or saddle-point argument.

A further difficulty is that the kernel block acquires its curvature through the dual variable γ. In the nominal two-player game, the KL regularizers already provide fixed curvature in the policy blocks. In the four-player robust game, by contrast, the p-block is strongly convex only with modulus proportional to γ. This is why the lower bound $\gamma_{\min} > 0$ and the condition $\tau_\nu \gamma_{\min} \geq 1/4$ enter the analysis: they ensure that the KL curvature in the kernel block is strong enough to dominate the mixed bilinear interaction between the follower policy and the adversarial kernel. Without such a condition, the grouped (π, γ)-vs.-(ν, p) game need not be convex-concave, and the standard optimistic mirror-descent analysis does not go through.

Our gradient-variation analysis controls four interacting blocks. In nominal NLHF, the unregularized part of each policy gradient depends only on the other policy; the KL term also depends on the policy being updated. Here the gradients additionally depend on the kernel and multiplier, so the variation of one block can depend on the movement of several other blocks. In particular, the p-gradient depends on the current policies and on the dual variable, while the γ-gradient is exactly the KL-budget slack of the current kernel. Establishing convergence therefore requires simultaneous control of the four variation terms. The proof absorbs their sum into the negative Bregman stability terms and separately bounds the surviving endpoint terms.

D Auxiliary Lemmas

We prove several auxiliary lemmas required for the theorems.

Lemma 1 (Universal upper bound on the dual multiplier)**.** *Assume $\rho > 0$ and $p^\star \in P_\delta$. Fix $(\pi, \nu) \in \Pi_\epsilon \times \Delta_m$, and define*

$$q_{\pi,\nu}(\gamma) := \min_{p \in P_\delta} \left\{ J(\pi, \nu; p) + \gamma\big(D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) - \rho\big) \right\}, \qquad \gamma \geq 0.$$

Then every maximizer of $q_{\pi,\nu}$ over $[0, \infty)$ lies in $[0, 1/\rho]$. In particular, one may choose the algorithmic upper bound $\gamma_{\max}$ so that $\gamma_{\max} \geq 1/\rho$.

Proof. Because $p^\star \in P_\delta$, for every $\gamma \geq 0$,

$$q_{\pi,\nu}(\gamma) \leq J(\pi, \nu; p^\star) - \rho\gamma.$$

At $\gamma = 0$,

$$q_{\pi,\nu}(0) = \min_{p \in P_\delta} J(\pi, \nu; p).$$

For fixed (π, ν), the only part of $J(\pi, \nu; p)$ that depends on p is the win-rate term, which always lies in $[0, 1]$. Hence

$$J(\pi, \nu; p^\star) - q_{\pi,\nu}(0) \leq 1.$$

Therefore, if $\gamma > 1/\rho$, then

$$q_{\pi,\nu}(\gamma) \leq J(\pi, \nu; p^\star) - \rho\gamma < J(\pi, \nu; p^\star) - 1 \leq q_{\pi,\nu}(0),$$

so such a γ cannot be optimal. The function $q_{\pi,\nu}$ is continuous, since it minimizes a continuous function over the compact set P_δ. It therefore attains its maximum on $[0, 1/\rho]$. □

Lemma 2 (Blockwise OMD inequalities)**.** *For a block $b \in \{\pi, \gamma, \nu, p\}$, let $s_\pi = s_\gamma = -1$ and $s_\nu = s_p = 1$. For a comparator u in that block's domain, define*

$$E_T^b(u) := -\frac{D_b(u, z_{T+1}^b)}{\eta} + s_b\langle g_0^b, z_1^b - u\rangle - s_b\langle g_T^b, z_{T+1}^b - u\rangle,$$

where z_t^b is the corresponding block of z_t, and the divergences D_b are those in Section 5. Algorithm 1 satisfies

$$\sum_{t=1}^T \langle g_t^\pi, \pi - \pi_t\rangle \le \frac{D_{\mathrm{KL}}(\pi\|\pi_1)}{\eta} + E_T^\pi(\pi) + \eta\sum_{t=1}^T \|g_t^\pi - g_{t-1}^\pi\|_\infty^2 - \frac{1}{2\eta}\sum_{t=1}^T D_{\mathrm{KL}}(\pi_{t+1}\|\pi_t), \tag{40}$$

$$\sum_{t=1}^T g_t^\gamma(\gamma - \gamma_t) \le \frac{(\gamma-\gamma_1)^2}{2\eta} + E_T^\gamma(\gamma) + \eta\sum_{t=1}^T |g_t^\gamma - g_{t-1}^\gamma|^2 - \frac{1}{4\eta}\sum_{t=1}^T |\gamma_{t+1} - \gamma_t|^2, \tag{41}$$

$$\sum_{t=1}^T \langle g_t^\nu, \nu_t - \nu\rangle \le \frac{D_{\mathrm{KL}}(\nu\|\nu_1)}{\eta} + E_T^\nu(\nu) + \eta\sum_{t=1}^T \|g_t^\nu - g_{t-1}^\nu\|_\infty^2 - \frac{1}{2\eta}\sum_{t=1}^T D_{\mathrm{KL}}(\nu_{t+1}\|\nu_t), \tag{42}$$

$$\sum_{t=1}^T \langle g_t^p, p_t - p\rangle \le \frac{D_{\mathrm{KL}}^{\mathrm{Bern}}(p\|p_1)}{\eta} + E_T^p(p) + \eta\sum_{t=1}^T \|g_t^p - g_{t-1}^p\|_2^2 - \frac{1}{2\eta}\sum_{t=1}^T D_{\mathrm{KL}}^{\mathrm{Bern}}(p_{t+1}\|p_t). \tag{43}$$

Proof. Apply Lemma 5 to each block with the signed gradient $s_b g_t^b$. The simplex entropy divergences satisfy $D_b(u, v) \ge \frac{1}{2}\|u-v\|_1^2$, the Euclidean divergence equals $\frac{1}{2}|u-v|^2$, and the Bernoulli divergence is at least $2\|u - v\|_2^2$. Thus every block satisfies the lemma's strong-convexity normalization. The corresponding dual norms are ℓ_∞, absolute value, ℓ_∞, and ℓ_2, respectively. Substitution gives the four inequalities, including their endpoint terms and the Euclidean factor $1/(4\eta)$. □

The constant-step OGDA proof requires a uniform bound on blockwise gradient variation.

Lemma 3 (Gradient-variation bounds)**.** *Let*

$$\pi_{\mathrm{ref,min}} := \min_{1\le i\le m} \pi_{\mathrm{ref},i}, \qquad L_\pi := \frac{\tau_\pi}{\epsilon\pi_{\mathrm{ref,min}}}, \qquad L_\nu := \frac{\tau_\nu}{\beta\pi_{\mathrm{ref,min}}}, \qquad L_p := \frac{\gamma_{\max}}{\delta(1-\delta)},$$

and

$$M_\delta := 2\log\frac{1-\delta}{\delta}.$$

For $t \ge 2$, let

$$z_t = (\pi_t, \gamma_t, \nu_t, p_t) \in X \times Y.$$

Then the block gradients satisfy

$$\|g_t^\pi - g_{t-1}^\pi\|_\infty \le L_\pi \|\pi_t - \pi_{t-1}\|_1 + \|\nu_t - \nu_{t-1}\|_1 + \|p_t - p_{t-1}\|_2, \tag{44}$$

$$|g_t^\gamma - g_{t-1}^\gamma| \le \sqrt{d}\, M_\delta \, \|p_t - p_{t-1}\|_2, \tag{45}$$

$$\|g_t^\nu - g_{t-1}^\nu\|_\infty \le \|\pi_t - \pi_{t-1}\|_1 + L_\nu \|\nu_t - \nu_{t-1}\|_1 + \|p_t - p_{t-1}\|_2, \tag{46}$$

$$\|g_t^p - g_{t-1}^p\|_2 \le \sqrt{d}\, \|\pi_t - \pi_{t-1}\|_1 + \sqrt{d}\, \|\nu_t - \nu_{t-1}\|_1 + L_p \|p_t - p_{t-1}\|_2 + \sqrt{d}\, M_\delta \, |\gamma_t - \gamma_{t-1}|. \tag{47}$$

Consequently, if we define for $t \ge 2$

$$S_t^\pi := \|\pi_t - \pi_{t-1}\|_1^2, \qquad S_t^\gamma := |\gamma_t - \gamma_{t-1}|^2, \qquad S_t^\nu := \|\nu_t - \nu_{t-1}\|_1^2, \qquad S_t^p := \|p_t - p_{t-1}\|_2^2,$$

and

$$L_\delta = L := \max\left\{1,\, L_\pi,\, L_\nu,\, L_p,\, \sqrt{d},\, \sqrt{d}\, M_\delta\right\},$$

then

$$\|g_t^\pi - g_{t-1}^\pi\|_\infty^2 \le 3L^2\big(S_t^\pi + S_t^\nu + S_t^p\big), \tag{48}$$

$$|g_t^\gamma - g_{t-1}^\gamma|^2 \le L^2 S_t^p, \tag{49}$$

$$\|g_t^\nu - g_{t-1}^\nu\|_\infty^2 \le 3L^2\big(S_t^\pi + S_t^\nu + S_t^p\big), \tag{50}$$

$$\|g_t^p - g_{t-1}^p\|_2^2 \le 4L^2\big(S_t^\pi + S_t^\gamma + S_t^\nu + S_t^p\big). \tag{51}$$

Proof. We prove the first-order bounds and then square them.

The π-block. From (55) of Lemma 6,

$$g^\pi(i) = \sum_{j=1}^m \nu_j P_p(i \succ j) - \tau_\pi \left(\log \frac{\pi_i}{\pi_{\mathrm{ref},i}} + 1\right).$$

For the win-rate component, since $\sum_j \nu_j = 1$,

$$\begin{aligned}
&\left|\sum_j \nu_{t,j} P_{p_t}(i \succ j) - \sum_j \nu_{t-1,j} P_{p_{t-1}}(i \succ j)\right| \\
&\quad\le \left|\sum_j (\nu_{t,j} - \nu_{t-1,j}) P_{p_t}(i \succ j)\right| + \left|\sum_j \nu_{t-1,j}\big(P_{p_t}(i \succ j) - P_{p_{t-1}}(i \succ j)\big)\right| \\
&\quad\le \|\nu_t - \nu_{t-1}\|_1 + \|p_t - p_{t-1}\|_\infty \le \|\nu_t - \nu_{t-1}\|_1 + \|p_t - p_{t-1}\|_2.
\end{aligned}$$

For the KL term, since every $\pi_t(i) \ge \epsilon \pi_{\mathrm{ref},i} \ge \epsilon \pi_{\mathrm{ref,min}}$,

$$\left|\log \frac{\pi_t(i)}{\pi_{\mathrm{ref},i}} - \log \frac{\pi_{t-1}(i)}{\pi_{\mathrm{ref},i}}\right| = |\log \pi_t(i) - \log \pi_{t-1}(i)| \le \frac{|\pi_t(i) - \pi_{t-1}(i)|}{\epsilon \pi_{\mathrm{ref,min}}}$$

by the mean-value theorem. Taking the maximum over i proves (44).

The γ-block. By (56) of Lemma 6,

$$g^\gamma = D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) - \rho.$$

For each coordinate $e = (i, j) \in E$, the derivative of the scalar Bernoulli-KL term with respect to p_e is

$$h_e'(u) = \log \frac{u(1 - p_e^\star)}{p_e^\star(1 - u)}.$$

Since both u and $p_e^\star$ lie in $[\delta, 1 - \delta]$, we have

$$|h_e'(u)| \le |\operatorname{logit}(u)| + |\operatorname{logit}(p_e^\star)| \le 2 \log \frac{1 - \delta}{\delta} = M_\delta.$$

Therefore, by the mean-value theorem and Cauchy–Schwarz,

$$|g_t^\gamma - g_{t-1}^\gamma| \le \sum_{e=1}^{d} M_\delta \, |p_{t,e} - p_{t-1,e}| \le \sqrt{d}\, M_\delta \, \|p_t - p_{t-1}\|_2,$$

which proves (45).

The ν-block. Use the same decomposition as for the π-block, with τ_ν in place of τ_π, the floor $\nu_{t,j} \ge \beta\pi_{\mathrm{ref,j}}$, and the transposed preference matrix. The logarithmic derivative is bounded by $1/(\beta\pi_{\mathrm{ref,min}})$, proving (46).

The p-block. From (58) of Lemma 6,

$$g_{ij}^p = w_{ij}(\pi,\nu) + \gamma\, h_{ij}(p_{ij}), \qquad h_{ij}(u) := \log \frac{u(1-p_{ij}^\star)}{p_{ij}^\star(1-u)}.$$

For the antisymmetric weights, fix $(i,j) \in E$. Then

$$\begin{aligned}|w_{ij}(\pi_t,\nu_t) - w_{ij}(\pi_{t-1},\nu_{t-1})| &= |\pi_{t,i}\nu_{t,j} - \pi_{t,j}\nu_{t,i} - \pi_{t-1,i}\nu_{t-1,j} + \pi_{t-1,j}\nu_{t-1,i}| \\ &\le |\pi_{t,i} - \pi_{t-1,i}| + |\pi_{t,j} - \pi_{t-1,j}| + |\nu_{t,i} - \nu_{t-1,i}| + |\nu_{t,j} - \nu_{t-1,j}| \\ &\le \|\pi_t - \pi_{t-1}\|_1 + \|\nu_t - \nu_{t-1}\|_1.\end{aligned}$$

Hence

$$\|w(\pi_t,\nu_t) - w(\pi_{t-1},\nu_{t-1})\|_2 \le \sqrt{d}\,\|\pi_t - \pi_{t-1}\|_1 + \sqrt{d}\,\|\nu_t - \nu_{t-1}\|_1.$$

For the KL-log term,

$$|h_{ij}'(u)| = \frac{1}{u(1-u)} \le \frac{1}{\delta(1-\delta)},$$

so

$$\|h(p_t) - h(p_{t-1})\|_2 \le \frac{1}{\delta(1-\delta)} \|p_t - p_{t-1}\|_2.$$

Also $|h_{ij}(u)| \le M_\delta$ on $[\delta, 1-\delta]$, so

$$\|h(p_{t-1})\|_2 \le \sqrt{d}\, M_\delta.$$

Therefore

$$\begin{aligned}\|g_t^p - g_{t-1}^p\|_2 &\le \|w(\pi_t,\nu_t) - w(\pi_{t-1},\nu_{t-1})\|_2 + \gamma_{\max}\|h(p_t) - h(p_{t-1})\|_2 + |\gamma_t - \gamma_{t-1}|\,\|h(p_{t-1})\|_2 \\ &\le \sqrt{d}\,\|\pi_t - \pi_{t-1}\|_1 + \sqrt{d}\,\|\nu_t - \nu_{t-1}\|_1 + L_p\|p_t - p_{t-1}\|_2 + \sqrt{d}\, M_\delta\,|\gamma_t - \gamma_{t-1}|,\end{aligned}$$

which proves (47).

Finally, (48)–(51) follow from

$$(a+b+c)^2 \le 3(a^2+b^2+c^2), \qquad (a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2).$$

□

Lemma 4 (Gradient variation is absorbed by the stability terms)**.** *Let*

$$V_T := \eta \sum_{t=1}^{T} \left(\|g_t^\pi - g_{t-1}^\pi\|_\infty^2 + |g_t^\gamma - g_{t-1}^\gamma|^2 + \|g_t^\nu - g_{t-1}^\nu\|_\infty^2 + \|g_t^p - g_{t-1}^p\|_2^2 \right),$$

and

$$\Xi_T := \frac{1}{2\eta} \sum_{t=1}^{T} \left(D_{\mathrm{KL}}(\pi_{t+1}\|\pi_t) + \tfrac{1}{2}|\gamma_{t+1} - \gamma_t|^2 + D_{\mathrm{KL}}(\nu_{t+1}\|\nu_t) + D_{\mathrm{KL}}^{\mathrm{Bern}}(p_{t+1}\|p_t) \right).$$

If

$$\eta \le \frac{1}{8L}, \tag{52}$$

where $L = L_\delta$ is the constant from Lemma 3 and $g_0^b = g_1^b$ for every block, then

$$V_T \le \Xi_T. \tag{53}$$

Proof. Because $g_0^\bullet = g_1^\bullet$, the $t = 1$ contribution to V_T is zero. Thus we sum from $t = 2$ onward. By Lemma 3,

$$V_T \le 11L^2\eta \sum_{t=2}^{T} \big(S_t^\pi + S_t^\gamma + S_t^\nu + S_t^p\big).$$

On the other hand, Pinsker's inequality and the Bernoulli Pinsker bound give

$$D_{\mathrm{KL}}(\pi_{t+1}\|\pi_t) \ge \frac{1}{2}\|\pi_t - \pi_{t+1}\|_1^2 = \frac{1}{2}S_{t+1}^\pi,$$
$$D_{\mathrm{KL}}(\nu_{t+1}\|\nu_t) \ge \frac{1}{2}\|\nu_t - \nu_{t+1}\|_1^2 = \frac{1}{2}S_{t+1}^\nu,$$
$$D_{\mathrm{KL}}^{\mathrm{Bern}}(p_{t+1}\|p_t) \ge 2\|p_t - p_{t+1}\|_2^2 = 2S_{t+1}^p.$$

Hence

$$\begin{aligned}\Xi_T &\ge \frac{1}{4\eta}\sum_{t=1}^{T} S_{t+1}^\pi + \frac{1}{4\eta}\sum_{t=1}^{T} S_{t+1}^\gamma + \frac{1}{4\eta}\sum_{t=1}^{T} S_{t+1}^\nu + \frac{1}{\eta}\sum_{t=1}^{T} S_{t+1}^p \\ &\ge \frac{1}{4\eta}\sum_{t=2}^{T+1} \big(S_t^\pi + S_t^\gamma + S_t^\nu + S_t^p\big) \\ &\ge \frac{1}{4\eta}\sum_{t=2}^{T} \big(S_t^\pi + S_t^\gamma + S_t^\nu + S_t^p\big).\end{aligned}$$

Therefore $V_T \le \Xi_T$ whenever

$$11L^2\eta \le \frac{1}{4\eta}, \qquad \text{i.e.} \qquad \eta \le \frac{1}{2\sqrt{11}\,L}.$$

The stronger condition (52) is sufficient. □

Lemma 5 (Optimistic mirror-descent bound with endpoint terms)**.** *Let Z be compact and convex, and let the differentiable mirror map h satisfy $D_h(v, w) \ge \frac{1}{2}\|v - w\|^2$ on Z. For $\eta > 0$, consider*

$$z_{t+1} = \arg\min_{z\in Z} \left\{ \langle 2g_t - g_{t-1}, z\rangle + \frac{1}{\eta} D_h(z, z_t) \right\}.$$

For every comparator $u \in Z$ and $T \ge 1$,

$$\begin{aligned}\sum_{t=1}^{T} \langle g_t, z_t - u\rangle \le{}& \frac{D_h(u, z_1) - D_h(u, z_{T+1})}{\eta} + \langle g_0, z_1 - u\rangle - \langle g_T, z_{T+1} - u\rangle \\ &+ \eta \sum_{t=1}^{T} \|g_t - g_{t-1}\|_*^2 - \frac{1}{2\eta}\sum_{t=1}^{T} D_h(z_{t+1}, z_t). \qquad (54)\end{aligned}$$

For maximization, apply this inequality to $-g_t$.

Proof. The update's first-order optimality condition and the three-point Bregman identity give

$$\eta\langle 2g_t - g_{t-1}, z_{t+1} - u\rangle \le D_h(u, z_t) - D_h(u, z_{t+1}) - D_h(z_{t+1}, z_t).$$

To evaluate the loss at z_t, use the exact identity

$$\begin{aligned}\langle g_t, z_t - u\rangle = \langle 2g_t - g_{t-1}, z_{t+1} - u\rangle &+ \langle g_{t-1}, z_t - u\rangle - \langle g_t, z_{t+1} - u\rangle \\ &- \langle g_t - g_{t-1}, z_{t+1} - z_t\rangle.\end{aligned}$$

Duality of the norms, Young's inequality, and the assumed normalization imply

$$\begin{aligned}-\langle g_t - g_{t-1}, z_{t+1} - z_t\rangle &\le \eta\|g_t - g_{t-1}\|_*^2 + \frac{\|z_{t+1} - z_t\|^2}{4\eta} \\ &\le \eta\|g_t - g_{t-1}\|_*^2 + \frac{D_h(z_{t+1}, z_t)}{2\eta}.\end{aligned}$$

Substitute these two relations into the first inequality and sum for $t = 1, \dots, T$. The Bregman differences telescope to $D_h(u, z_1) - D_h(u, z_{T+1})$. The inner-product differences telescope to $\langle g_0, z_1 - u\rangle - \langle g_T, z_{T+1} - u\rangle$. The remaining stability coefficient is $-1/(2\eta)$, which gives (54). □

Lemma 6 (Partial gradients)**.** *For $z = (\pi, \gamma, \nu, p) \in X \times Y$, the block gradients of F are:*

$$g^{\pi}(i) = \sum_{j=1}^{m} \nu_j \, P_p(i \succ j) - \tau_\pi \left(\log \frac{\pi_i}{\pi_{\mathrm{ref},i}} + 1 \right), \qquad i = 1, \dots, m, \tag{55}$$

$$g^{\gamma} = D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) - \rho, \tag{56}$$

$$g^{\nu}(j) = \sum_{i=1}^{m} \pi_i \, P_p(i \succ j) + \tau_\nu \left(\log \frac{\nu_j}{\pi_{\mathrm{ref},j}} + 1 \right), \qquad j = 1, \dots, m, \tag{57}$$

$$g^{p}_{ij} = w_{ij}(\pi, \nu) + \gamma \log \frac{p_{ij}(1 - p^\star_{ij})}{p^\star_{ij}(1 - p_{ij})}, \qquad (i, j) \in E. \tag{58}$$

Proof. Differentiate (17). The π- and ν-gradients are obtained from the linear win-rate term and the KL regularizers. The γ-gradient is immediate because F is affine in γ. For (58), the coefficient of p_{ij} in the win-rate decomposition is $w_{ij}(\pi, \nu)$, and

$$\frac{\partial}{\partial p_{ij}} D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) = \log \frac{p_{ij}(1 - p^\star_{ij})}{p^\star_{ij}(1 - p_{ij})}. \qquad \square$$

E IMDb Continuation Task: Experiment Details

E.1 Nominal Preference and Annotator Groups

We use the IMDb review corpus [23] as a source of continuation prompts. Each prompt is the first sentence of a review; the task is to continue it. We sample $N = 8{,}000$ prompts and generate $m = 8$ candidate continuations per prompt using `google/flan-t5-large` [11] with diverse instruction templates and sampling temperatures. The full set of prompt-level preference kernels is pre-computed and cached so that NashMD and all robust variants are trained and evaluated on identical finite action sets.

To model heterogeneous human preferences we construct three annotator groups: **Group A** (comprehensive/analytical) prefers longer, more detailed continuations with stronger evidence of analysis; **Group B** (efficient/concise) prefers short, direct, and decisive continuations; and **Group C** (balanced/nuanced) prefers responses that acknowledge both strengths and weaknesses. For prompt x and group $g \in \{A, B, C\}$, each candidate i receives a scalar score $s_i^{(g)}$. These scores are converted into a group-specific Bradley–Terry pairwise preference kernel:

$$P_{ij}^{(g)} = \sigma\big(\beta \, (s_i^{(g)} - s_j^{(g)})\big), \quad i < j, \tag{59}$$

with $P_{ji}^{(g)} = 1 - P_{ij}^{(g)}$ and $P_{ii}^{(g)} = \frac{1}{2}$. Writing $p^{(g)} \in [0, 1]^d$, $d = \binom{m}{2} = 28$, for the upper-triangular vector, the nominal preference kernel is $p^\star = \sum_{g \in \{A,B,C\}} \alpha_g \, p^{(g)}$ with fixed mixture weights α. Perturbing the mixture weights away from α via a Dirichlet draw produces *group-shift* kernels that lie in the convex hull of the group kernels and serve as a model of structured population-level preference shift (Section E.4).

E.2 Optimization Algorithms

NashMD baseline. The NashMD baseline solves the nominal KL-regularized game

$$\max_{\pi \in \Delta^m} \min_{\nu \in \Delta^m} J(\pi, \nu; P^\star) = \pi^\top P^\star \nu - \tau \, D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}}) + \tau \, D_{\mathrm{KL}}(\nu \| \pi_{\mathrm{ref}}) \tag{60}$$

with nominal kernel $P^\star = P(p^\star)$ fixed throughout training. We use simultaneous Optimistic Gradient Descent-Ascent (OGDA) on the policy pair (π, ν) with step size $\eta_{\mathrm{Nash}} = 0.30$.

Four-player robust algorithm. The robust variant hedges against kernel misspecification within the Bernoulli-KL ball $\mathcal{U}_\rho = \{p : D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) \le \rho\}$ by solving

$$\max_{\pi \in \Delta^m} \min_{\substack{p : D_{\mathrm{KL}}^{\mathrm{Bern}}(p \| p^\star) \le \rho \\ p \in [\delta, 1-\delta]^d}} \min_{\nu \in \Delta^m} J(\pi, \nu; P(p)). \tag{61}$$

The inner minimization is solved approximately by alternating between the closed-form follower best-response and the kernel update $p_{ij} \leftarrow \sigma\big(\mathrm{logit}(p^\star_{ij}) - w_{ij}(\pi,\nu)/\max(\gamma,\gamma_{\min})\big)$ for $K = 8$ inner steps, then the outer (π,γ) update is performed with OMD at step sizes $(\eta_\pi,\eta_\gamma) = (0.25, 0.15)$.

E.3 Hyperparameters

Table 2: Hyperparameters for the IMDb continuation experiments.

Component	**Value**
Candidates per prompt m	8
Kernel dimension d	$\binom{8}{2} = 28$
Number of prompts N	8,000
Generator model	`flan-t5-large`
KL temperature τ	$\{0.001, 0.005, 0.01, 0.05, 0.1\}$
NashMD step size η_{Nash}	0.30
Robust step sizes (η_π,η_γ)	(0.25, 0.15)
Inner steps K	8
Training radii ρ_{train}	{0.005, 0.01, 0.02, 0.05, 0.10, 0.15}
Evaluation radii ρ_{eval}	{0.05, 0.10, 0.15, 0.20, 0.30, 0.50, 0.75, 1.0}
Kernel clipping δ	10^{-3}
Dual range $[\gamma_{\min},\gamma_{\max}]$	[0.02, 500]

E.4 Results

Given the robust policy π_R and the NashMD baseline π_N, we measure the win rate using 25. The sampled kernels for evaluation are (i) the KL-ball boundary $\mathcal{U}_{\rho_{\mathrm{eval}}}$ sampled uniformly at random, used for Table 3, and (ii) a Dirichlet-perturbed annotator mixture $\tilde{p} = \sum_g \tilde{\alpha}_g p^{(g)}$, $\tilde{\alpha} \sim \mathrm{Dirichlet}(\kappa\alpha)$, projected to the KL-ball boundary at radius ρ_{eval}, used in Table 4. Annotator group-shift kernels remain in the low-dimensional mixture subspace, making them a more structured perturbation family.

Table 3: Win rate against NashMD under random kernel perturbations, averaged over $N = 8000$ IMDb prompts with $m = 8$ candidates. Each entry reports $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ from (25). Values above 0.50 indicate the robust policy is preferred under the majority of sampled kernels.

	Evaluation radius ρ_{eval}				
ρ_{train}	0.05	0.10	0.20	0.30	0.50
0.01	0.986	0.967	0.927	0.894	0.844
0.02	0.985	0.965	0.925	0.891	0.841
0.05	0.984	0.961	0.918	0.883	0.832
0.10	0.978	0.949	0.901	0.866	0.814
0.15	0.934	0.894	0.844	0.810	0.766

Table 4: $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ under group-shift kernels at $\tau = 0.01$, averaged over 8,000 prompts. Bold marks the highest value in each column.

ρ_{train}	$\rho_{\mathrm{eval}} = 0.05$	0.10	0.20	0.30	0.50
0.005	**0.974**	**0.945**	**0.897**	**0.863**	**0.818**
0.01	0.974	0.944	0.897	0.862	0.818
0.02	0.973	0.943	0.896	0.862	0.818
0.05	0.970	0.939	0.893	0.859	0.816
0.10	0.962	0.928	0.884	0.851	0.811
0.15	0.917	0.882	0.840	0.811	0.775

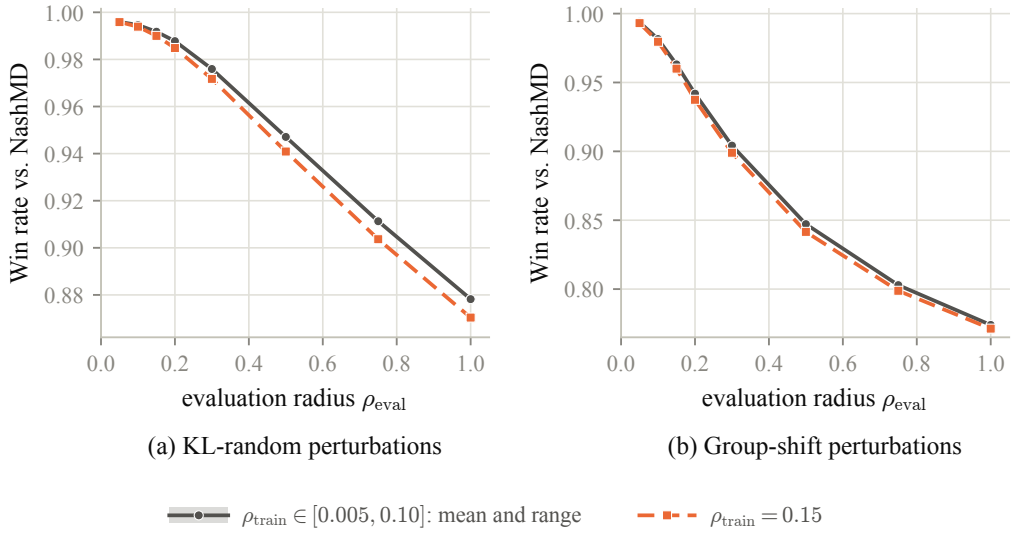
Table 5 reports $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ at the best training radius for each KL-regularization temperature $\tau \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$. The strongest overall regime is $\tau = 0.05$: the robust policy wins with probability above 97% for $\rho_{\mathrm{eval}} \le 0.30$. At very small τ the NashMD baseline is

already highly concentrated, reducing the marginal gain from robust training; at large τ both policies approach the uniform distribution and the adversary has limited room to exploit kernel perturbations.

Table 5: $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ at the best ρ_{train} for each τ (KL-random kernels).

τ	best ρ_{tr}	$\rho_{\mathrm{eval}} = 0.05$	0.10	0.20	0.30	0.50	1.0
0.001	0.005	0.972	0.944	0.896	0.859	0.810	0.735
0.005	0.005	0.980	0.956	0.911	0.877	0.826	0.750
0.01	0.005	0.986	0.967	0.929	0.896	0.847	0.768
0.05	0.10	**0.996**	**0.995**	**0.988**	**0.977**	**0.948**	**0.879**
0.1	0.10	0.945	0.931	0.903	0.880	0.841	0.775

Figure 2 shows the full win-rate curves at $\tau = 0.05$. Training radii $\rho_{\mathrm{train}} \le 0.10$ are statistically indistinguishable from one another at every evaluation radius (their pairwise differences never exceed the 95% confidence half-width), while $\rho_{\mathrm{train}} = 0.15$ is significantly worse, with the gap widening as ρ_{eval} grows.



(a) KL-random perturbations

(b) Group-shift perturbations

Figure 2: $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ at $\tau = 0.05$, for each training radius $\rho_{\mathrm{train}} \in \{0.005, 0.01, 0.02, 0.05, 0.10, 0.15\}$. Radii $\rho_{\mathrm{train}} \in [0.005, 0.10]$ (gray band and mean) are statistically tied at every evaluation radius; $\rho_{\mathrm{train}} = 0.15$ (orange) is significantly worse, and increasingly so as ρ_{eval} grows.

Table 6 gives the full $(\tau, \rho_{\mathrm{train}})$ interaction at $\rho_{\mathrm{eval}} = 0.30$. For $\tau \le 0.01$ the smallest training radius is uniformly best. For $\tau \ge 0.05$ the optimum shifts to $\rho_{\mathrm{train}} = 0.10$, though the improvement over smaller radii is modest.

Table 6: $\mathrm{WinRate}(\pi_R \text{ vs } \pi_N;\ \rho_{\mathrm{eval}})$ at $\rho_{\mathrm{eval}} = 0.30$, by $(\tau, \rho_{\mathrm{train}})$ (KL-random kernels). Bold marks the row-wise maximum.

$\tau \setminus \rho_{\mathrm{tr}}$	0.005	0.01	0.02	0.05	0.10	0.15
0.001	**0.860**	0.857	0.850	0.832	0.813	0.780
0.005	**0.877**	0.874	0.870	0.857	0.831	0.784
0.01	**0.896**	0.895	0.893	0.883	0.866	0.810
0.05	0.976	0.976	0.976	0.976	**0.977**	0.972
0.1	0.859	0.860	0.862	0.868	**0.880**	0.717

Algorithm 2 Sampled four-player step

Require: leader π_θ, follower $\nu_{\theta'}$ (separate LoRA adapters, both initialized at π_{ref}); head g_ψ (zero-initialized output layer); multiplier $\gamma \geq 0$; baselines b_L, b_F; τ, ρ, $\gamma_{\max}$, step sizes $\eta_\theta, \eta_{\theta'}, \eta_\gamma$, batch size B, head steps K, clip level c, EMA weight λ

for each iteration **do**
  draw prompts $x_1, \ldots, x_B$; sample $y_L^i \sim \pi_\theta(\cdot \mid x_i)$ and $y_F^i \sim \nu_{\theta'}(\cdot \mid x_i)$
  $q_i \leftarrow P^\star(y_L^i \succ y_F^i \mid x_i)$ (BT judge or PairRM); $\kappa_L^i \leftarrow \mathrm{clip}\big(\log \frac{\pi_\theta(y_L^i|x_i)}{\pi_{\mathrm{ref}}(y_L^i|x_i)}, \pm c\big)$, κ_F^i likewise
  $p_i \leftarrow p_i(\psi) = \sigma\big(\mathrm{logit}\, q_i + \delta_\psi(y_L^i, y_F^i)\big)$ ▷ head as at the start of the step
  $R_i^L \leftarrow p_i - \tau\kappa_L^i$, $R_i^F \leftarrow -p_i - \tau\kappa_F^i$ ▷ p_i held constant
  $b_L \leftarrow \lambda b_L + (1-\lambda)\frac{1}{B}\sum_i R_i^L$, $b_F \leftarrow \lambda b_F + (1-\lambda)\frac{1}{B}\sum_i R_i^F$
  $\hat{g}_\theta \leftarrow \frac{1}{B}\sum_i (R_i^L - b_L)\nabla_\theta \log \pi_\theta(y_L^i \mid x_i)$, $\hat{g}_{\theta'} \leftarrow \frac{1}{B}\sum_i (R_i^F - b_F)\nabla_{\theta'} \log \nu_{\theta'}(y_F^i \mid x_i)$
  for $k = 1, \ldots, K$ **do** ▷ same batch, no new samples
    gradient step on ψ (descent) for $\mathcal{L}_\psi = \frac{1}{B}\sum_i p_i(\psi) + \gamma\big(\frac{1}{B}\sum_i d_{\mathrm{Bern}}(p_i(\psi)\|q_i) - \rho\big)$
  end for
  $\gamma \leftarrow \mathrm{clip}\,\big(\gamma + \eta_\gamma(\frac{1}{B}\sum_i d_{\mathrm{Bern}}(p_i(\psi)\|q_i) - \rho),\, 0,\, \gamma_{\max}\big)$ with ψ after the head steps
  $\theta \leftarrow \theta + \eta_\theta \hat{g}_\theta$, $\theta' \leftarrow \theta' + \eta_{\theta'}\hat{g}_{\theta'}$ ▷ Adam, gradient-norm clipping
end for

F TL;DR Summarization Task: Experiment Details

F.1 From Algorithm 1 to the sampled implementation

For a language model the objects of Algorithm 1, sums over m responses, simplex mirror steps and the exact kernel tilt, are not available, because a response is a token sequence. We follow the route of Munos et al. [29] (their Appendix F): sample a pair instead of summing, replace each proximal step by one stochastic policy-gradient step, and keep the reference KL inside the reward as a log-ratio, which is the score-function form of $\nabla_\theta \mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}})$ because $\mathbb{E}\,\nabla_\theta \log \pi_\theta = 0$. Their opponent, a token-level geometric mixture of the policy and the reference, is replaced here by the trained follower $\nu_{\theta'}$: at the nominal equilibrium the follower's best response is the leader itself, so a mixture of leader and reference is a good stand-in, whereas under a perturbed kernel the best response is a different policy. We write $\pi_\theta, \nu_{\theta'}$ for the leader and follower that Appendices F.4–F.5 call $\pi_{\mathrm{L}}, \pi_{\mathrm{F}}$, and $d_{\mathrm{Bern}}(a\|b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ for the Bernoulli KL of a single comparison.

What the sampled method approximates. *(i) Noise:* each update uses one comparison per prompt and a small batch, so it is a noisy estimate of the corresponding block gradient of the tabular game. The mirror steps of Algorithm 1 move the logits additively, whereas a gradient step on the adapter parameters weights each response by its probability. *(ii) Baselines:* the leader and the follower subtract exponential-moving-average baselines, which need no value function. *(iii) KL proxy:* κ is a clipped sequence-level log-ratio, which biases the KL term whenever the clip is active. *(iv) Optimization:* a few gradient steps of the head stand in for the exact kernel tilt; there is no optimism and no averaging; γ is clipped to $[0, \gamma_{\max}]$, so Assumption 2-(3) and (4) need not hold. *(v) Start:* both adapters start at π_{ref}, so the two policies coincide at the start and, in expectation, the kernel player has no incentive to move until they separate; early training behaves like the nominal algorithm.

F.2 Fine-Tuning Models.

We initialize the models from `Qwen2.5-1.5B-Instruct` [51] and `Gemma-2-2B` [44] as both the initial and the reference model across the respective experiments and fine-tune with LoRA [19] for 5000 gradient steps with KL regularization $\tau = 0.005$. The four-player game formulation was compared with adversarial budget $\rho \in \{0.01, 0.05\}$. The leader and follower policies carry separate LoRA adapters attached to the same frozen backbone.

F.3 Nominal Preference Proxy

We train a Bradley-Terry(BT) preference judge model to serve as the nominal preference proxy. The judge is a scalar score model $s_\phi : (x, y) \to \mathbb{R}$ built on `DistilBERT-base-uncase` [39] backbone

with a linear scalar head, and fine-tuned with the standard BT cross-entropy loss on the TL;DR dataset:

$$\mathcal{L}_{\mathrm{BT}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)}\left[\log \sigma\big(s_\phi(x,y_w) - s_\phi(x,y_l)\big)\right], \tag{62}$$

where (y_w, y_l) denotes the preferred and dispreferred responses respectively. Training uses 83,573 comparison pairs. After 3 epochs the judge achieves 72% validation accuracy. During training with both NashMD and Four-Player the pairwise preference is calculated as:

$$P^*(y_a \succ y_b \mid x) = \sigma\big(s_\phi(x,y_a) - s_\phi(x,y_b)\big) \tag{63}$$

F.4 Adversary Proxy

The kernel-adversary in the four-player game is parameterized by an *antisymmetric* perturbation head $\delta_\psi(y_a, y_b)$ of the nominal preference logit. The head computes

$$\delta_\psi(y_a,y_b) \;=\; g_\psi(u_{ab}) - g_\psi(u_{ba}), \qquad g_\psi : \mathbb{R}^7 \to \mathbb{R}, \tag{64}$$

where g_ψ is a two-layer MLP: $\mathbb{R}^7 \xrightarrow{W_1} \mathbb{R}^{64} \xrightarrow{\tanh} \mathbb{R}^{64} \xrightarrow{W_2} \mathbb{R}$, with the output layer W_2 initialized to zero so that $\delta_\psi(y_a, y_b) = 0$ (no perturbation) at initialization. The algorithm therefore initializes at the nominal preference and gradually grows the adversarial perturbation. Antisymmetry in (64) ensures $\delta_\psi(y_a, y_b) = -\delta_\psi(y_b, y_a)$ and hence $P_\psi(y_a \succ y_b) + P_\psi(y_b \succ y_a) = 1$ at all iterates.

The perturbed preference is

$$P_\psi(y_a \succ y_b \mid x) \;=\; \sigma\big(\operatorname{logit} P^*(y_a \succ y_b \mid x) + \delta_\psi(y_a,y_b)\big). \tag{65}$$

The Bernoulli-KL ambiguity constraint $\mathbb{E}\big[D_{\mathrm{KL}}^{\mathrm{Bern}}(p_\psi \,\|\, p^\star)\big] \le \rho$ is enforced softly through the dual variable γ in the Lagrangian payoff. We additionally clip the predicted logit shift before applying the sigmoid in 65 to prevent P_ψ from saturating to $\{0, 1\}$ to stabilize the dual ascent.

F.5 Losses.

For a mini-batch with samples $y_L \sim \pi_{\mathrm{L}}$ and $y_F \sim \pi_{\mathrm{F}}$, we define $p_\psi = P_\psi(y_L \succ y_F \mid x)$ and KL proxies $\kappa_L = \mathrm{clip}(\log \pi_{\mathrm{L}}(y_L)/\pi_{\mathrm{ref}}(y_L), -c, c)$, $\kappa_F = \mathrm{clip}(\log \pi_{\mathrm{F}}(y_F)/\pi_{\mathrm{ref}}(y_F), -c, c)$. The three player losses are

$$\mathcal{L}_\psi = \mathbb{E}\big[p_\psi\big] + \gamma\left(\mathbb{E}[D_{\mathrm{KL}}^{Bern}(p_\psi \,\|\, P^*)] - \rho\right), \tag{66}$$

$$\mathcal{L}_\pi = -\mathbb{E}\big[(p_\psi - \tau\kappa_L - b_L)\,\log \pi_{\mathrm{L}}(y_L)\big], \tag{67}$$

$$\mathcal{L}_\nu = -\mathbb{E}\big[(-p_\psi - \tau\kappa_F - b_F)\,\log \pi_{\mathrm{F}}(y_F)\big], \tag{68}$$

where b_L, b_F are EMA baselines for variance reduction, and the adversary ψ *minimizes* (66) (gradient descent) while the dual variable is updated by clipped projected gradient ascent:

$$\gamma \;\leftarrow\; \mathrm{clip}\Big(\gamma + \eta_\gamma\left(\mathbb{E}[D_{\mathrm{KL}}^{Bern}] - \rho\right),\; 0,\; \gamma_{\max}\Big). \tag{69}$$

Each outer step interleaves K adversary updates per (π_L, π_F, γ) update. We default to $K = 3$ as we obseved $K = 1$ is too weak (the constraint is far from active) while $K \ge 3$ tracks the constraint budget closely.

F.6 Evaluation with LLM Judges

We employ three comparatively larger open-source models—DeepSeek-R1-Distill-Qwen-32B [12], Gemma-4-31B [44], and Nemotron-3-Nano-Omni [32]—to obtain preference signals querying with the following prompt template:

```
You are an expert judge evaluating the quality of two generated
summaries.
Post to summarize: ⟨text⟩
Summary A: ⟨summary1⟩
Summary B: ⟨summary2⟩
Evaluate which summary is better based on: (1) Accuracy, (2)
Coherence, (3) Conciseness, (4) Helpfulness.
```

where $\langle$text$\rangle$ corresponds to the Reddit post x, $\langle$summary1$\rangle$ to the policy response $y \sim \pi(\cdot \mid x)$, and $\langle$summary2$\rangle$ to the baseline response $y' \sim \pi_{\text{base}}(\cdot \mid x)$. We estimate the preference probability $P_j^{\text{ext}}(\pi \succ \pi_{\text{base}})$ as the fraction of comparisons in which the policy output is preferred over 1000 prompts from the TL;DR dataset. Following Munos et al. [29], each comparison is treated as a Bernoulli trial and uncertainty is quantified using Clopper–Pearson 95% confidence intervals; for $n = 1000$, the worst-case half-width is approximately ± 0.032. Four player policies preserve strong nominal performance against the NashMD baseline. Training against the preference uncertainty has not collapsed into overly conservative solutions, rather, the robust policies still improve the nominal pairwise objective with the pairwise judges across evaluation criteria.

F.7 Non-BT Nominal Kernel: Construction and Diagnostic

The TL;DR experiment in Section 6.3 uses a non-Bradley–Terry (non-BT) nominal kernel rather than the standard, reproducible Bradley–Terry (BT) construction (BT judge, Appendix F.3) [41], which is antisymmetric by design: $P^\star(y_a \succ y_b \mid x) = \sigma\big(s_\phi(x, y_a) - s_\phi(x, y_b)\big)$. The ambiguity set $\mathcal{P}$ in Eq. (8) is a Bernoulli-KL ball over the $d = \binom{m}{2}$ pairwise probabilities themselves (imposed in expectation over sampled pairs in TL;DR, Appendix F.5), with no constraint of the form $\operatorname{logit}(p_{ij}) = s_i - s_j$: even from a BT center, the worst-case kernel found by the adversary may already be cyclic or otherwise not representable by any scalar reward. We go further and construct a nominal kernel that is itself non-BT, described next, together with a diagnostic confirming it is measurably so.

Non-BT nominal kernel. We replace the BT judge with a directly parameterized pairwise kernel,

$$P_\phi^{\text{pair}}(y_a \succ y_b \mid x) = \sigma\big(L^{\text{pair}}(x, y_a, y_b)\big), \qquad L^{\text{pair}}(x, y_a, y_b) = \tfrac{c}{2}\big[q_\phi(x, y_a, y_b) - q_\phi(x, y_b, y_a)\big],$$

where q_ϕ is a single joint forward pass over the post and both candidate summaries, instantiated with PairRM [21] (a 0.4B DeBERTa-v3-large cross-encoder). Each candidate's representation is conditioned on its opponent, so its score is not a function of (x, y) alone and the kernel is not constrained to be BT. Because q_ϕ is order-dependent, the difference of the two orderings is what gives $L^{\text{pair}}(a, b) = -L^{\text{pair}}(b, a)$, hence $P^{\text{pair}}(a \succ b) + P^{\text{pair}}(b \succ a) = 1$ exactly. Here $c > 0$ is a scalar temperature fit by maximum likelihood on TL;DR human comparisons. The policy is Qwen2.5-1.5B-Instruct. The adversary head, Bernoulli-KL constraint, dual update and NashMD baseline are as in Appendix F.4–F.5; the only change is that three of the adversary's seven input features, which used per-response rewards, now carry the pairwise logit.

The kernel is not BT. Let $L_{ij} = \operatorname{logit} P_\phi^{\text{pair}}(y_i \succ y_j \mid x)$. A BT kernel has $L_{ij} = s_i - s_j$, so for every triple $C := L_{ij} + L_{jk} + L_{ki} = 0$: the scores telescope away, for any reward model. The least-squares BT fit to a triple removes $C/3$ from each edge, so $|C|/3$ is the per-edge distance to the nearest BT kernel. On 1,000 triples of real model outputs (per prompt, one summary from each of NashMD-PG and the two robust policies trained under this kernel), the mean best-BT projection error is 0.384 logits per edge (3.10 percentage points in preference probability), the median $|C|$ is 0.96, and 1.4% of triples are outright cyclic. Edges are confident ($|L| \approx 3$ logits on average), so a 0.38-logit residual rarely flips a sign: cycles are rare, yet the departure from BT is present in nearly every triple. The nominal kernel used in training is therefore measurably non-BT.

F.8 Additional Results with a Bradley-Terry Nominal Kernel

For completeness, we also report results under the standard Bradley–Terry (BT) nominal kernel (Appendix F.3) that Section 6.3's non-BT experiment departs from. Table 7 reports the win rate of the robust policy against NashMD on the same protocol as Table 1 (three external judges, 1000 TL;DR prompts, Clopper–Pearson half-width $\approx \pm 0.032$ at $n = 1000$), and additionally covers a second fine-tuning backbone, Gemma-2B, alongside Qwen-1.5B.

Table 7: Win rate against NashMD under external LLM judges, evaluated on 1000 TL;DR prompts, with a Bradley-Terry nominal kernel (cf. Table 1, which uses a non-BT kernel).

Fine-Tuning Model	ρ_{train}	Win Rate under Evaluation Model		
		DeepSeek-R1	Gemma-4	Nemotron-3
Qwen-1.5B	0.01	0.496	**0.506**	**0.764**
	0.05	**0.574**	**0.593**	**0.568**
Gemma-2B	0.01	**0.542**	**0.518**	0.462
	0.05	**0.640**	**0.720**	**0.583**

G LLM Judge Reasoning Examples

Example 1 (`prompt_id` **57**)

Disagreement: Nemotron-3-Nano → **A**; Gemma-4-31B → **B**; DeepSeek-R1-32B → **B**

Post (r/relationships)

Title: Me [20m] just got out of an 8 month relationship with my gf [19F]. I am very emotional, she wasn't.

I am far from the stereotypical macho guy. I love romance [. . .] She never speaks about her feelings, never told me she loved me without me having to say it first [. . .] My question aimed at both genders: Are girls generally emotional, do you ladies like to speak about emotions? Do you generally like when the guy is cute with you and calls you babe, and says he loves you? [. . .]

Summary A (NashMD-PG)

OP is seeking advice on whether girls generally like to be romantic and express emotions, and if they prefer a partner who is affectionate and expresses love.

Summary B (Four-Player ρ=0.05)

The OP is seeking to understand if girls generally express emotions and appreciate romantic displays of affection, as he is a romantic guy who enjoys being cute and expressing his love. He's questioning if this is something most girls find appealing.

Nemotron-3-Nano (verdict: **A**, 14.4 s)

[Conciseness] A is more concise (fewer words). B includes extra info about his own traits—possible unnecessary padding. So A wins on conciseness. [. . .]
[Helpfulness] A directly states the request for advice. B adds context that may help reader understand OP's perspective, but the key point—seeking advice about girls' preferences—is stated more succinctly in A. [. . .]
Overall, Summary A seems better due to higher conciseness and directness while still accurate.

Gemma-4-31B (verdict: **B**, 67.3 s)

[Helpfulness] Summary B captures the "why" of the post whereas A only captures the "what". The OP is not asking a generic question about gender; he is questioning his own experience in light of his personality and recent breakup. This makes B a more complete reflection of the source text. [. . .]
Summary A is more concise; Summary B is more contextually complete without becoming wordy. The motivation behind the question is a key part of the post, so B is slightly better overall.

DeepSeek-R1-Distill-Qwen-32B (verdict: **B**, 58.6 s)

[Helpfulness] Summary A captures the main question but lacks some context. Summary B provides more background, making it clearer why OP is asking this and what he has experienced. [. . .]
Summary B includes more details about OP's romantic nature and his girlfriend's behaviour, which are crucial for understanding his perspective.

Example 2 (`prompt_id` 125)

Disagreement: Nemotron-3-Nano → **A**; Gemma-4-31B → **B**; DeepSeek-R1-32B → **B**

Post (r/dating_advice)

Title: I (21,M) need help getting a girl (22) who used to like me interested again.

I recently started messaging a girl I knew from high school and it seemed to be going somewhere, but a couple of bad decisions and me not being honest enough in my replies have led me into a rut. She seems to have lost interest and only replies when I message first. How do I get from that to a position where I can ask her out? [. . .]

Summary A (NashMD-PG)

OP (21M) is trying to rekindle a past connection with a 22F he knows from high school. He's lost some mojo and is unsure how to move from a passive messaging pattern to asking her out.

Summary B (Four-Player ρ=0.05)

A 21-year-old man is trying to rekindle a connection with a 22-year-old woman he knew from high school, but after some miscommunications, she's lost interest. He's seeking advice on how to get back in her good graces and ask her out.

Nemotron-3-Nano (verdict: **A**, 16.9 s)

[Helpfulness] Summary A mentions the "passive messaging pattern", capturing the specific current dynamic (she only replies when prompted). Summary B says "lost interest" but does not identify this nuance. [. . .]
Summary A captures the specific dynamic slightly better; final verdict A.

Gemma-4-31B (verdict: **B**, 74.8 s)

[Accuracy] Summary B captures the cause-effect relationship: bad decisions → lost interest → needs advice. Summary A's "lost some mojo" reflects a side detail from the Fun Facts section rather than the primary driver of the conflict in the main text. [. . .]
[Helpfulness] Summary B provides a more complete narrative arc (the attempt, the mistake, the result, the goal), giving better context for advice.

DeepSeek-R1-Distill-Qwen-32B (verdict: **B**, 161.7 s)

[Accuracy] Summary B includes specific reasons for her losing interest (miscommunications / bad decisions), which aligns with the post. Summary A is vague, saying only he "lost some mojo", which does not capture the cause of her disengagement. [. . .]
[Helpfulness] B explains *why* she lost interest, providing context that would help readers offer more targeted advice.

Example 3 (`prompt_id` 114)

Disagreement: Nemotron-3-Nano → **B**; Gemma-4-31B → **B**; DeepSeek-R1-32B → **A**

Post (r/relationships)

Title: [m24] Arranged a date and it was postponed. Am I wasting my time?

A few weeks ago I met a girl, got her number and arranged to meet up. The day before, she asked to postpone because of an exam. I said sure; she then said she had exams for another couple of weeks. After they finished, I texted her on Saturday night [...] and got the response "no but [...]" Am I wasting my time?

Summary A (NashMD-PG)

OP is confused and questioning if he's wasting time after postponing a date with a girl who initially agreed to meet up but then said she'd be busy for a few more weeks.

Summary B (Four-Player ρ=0.05)

The OP arranged a date with a girl but she kept postponing it due to exams, leading him to question if he wasted his time. He's confused by her lack of clear communication and is wondering if she's not interested.

Nemotron-3-Nano (verdict: **B**, 12.0 s)

[Accuracy] Summary A says she "initially agreed" and was "busy for a few more weeks"—neither is quite right. She did not explicitly agree before the postponement, and the exams are now finished. Summary B accurately states she postponed *due to exams* and captures his uncertainty about her interest. [...]
[Helpfulness] B conveys both the postponement reason and his underlying concern (is she interested?), which is the real question driving the post.

Gemma-4-31B (verdict: **B**, 80.1 s)

[Accuracy] Summary A says "after postponing a date"—grammatically ambiguous, implying OP did the postponing, whereas the girl was the one who postponed. Summary B correctly identifies that *she* kept postponing, and also captures his underlying concern about her interest level. [...]
[Helpfulness] B explicitly states the core of OP's dilemma: whether she is actually interested. Summary A misses this psychological dimension entirely.

DeepSeek-R1-Distill-Qwen-32B (verdict: **A**, 70.8 s)

[Accuracy] Summary A captures the postponement and his confusion closely. Summary B adds that he is confused by "a lack of clear communication", which is an interpretation not explicitly stated in the post—the OP is confused about whether she is interested, not about communication clarity per se. [...]
[Conciseness] Summary A sticks strictly to the facts from the post. Summary B introduces an extra interpretive layer which might be considered redundant.

Example 4 (`prompt_id` 71)

Disagreement: Nemotron-3-Nano → **B**; Gemma-4-31B → **B**; DeepSeek-R1-32B → **A**

Post (r/relationships)

Title: I [F21] feel frustrated with my boyfriend's [M24] laziness. Am I justified?

I have a well-paying part-time job and am a full-time student; he works full time but doesn't make ends meet and I often cover his expenses. He plays video games while I'm trying to work on important assignments. When I asked him for help with housework he got mad, left without saying goodbye, and said *we should just break up* [...] Am I justified in feeling frustrated?

Summary A (NashMD-PG)

OP (F21) is frustrated with her boyfriend's (M24) lack of help and appreciation, especially during her stressful school deadlines. She feels unappreciated and **wants to break up** because he doesn't contribute to household chores or offer support.

Summary B (Four-Player ρ=0.05)

The poster is frustrated with her boyfriend's lack of effort in contributing to household responsibilities and feeling unappreciated despite her hard work and stress. She feels he's only concerned with his own leisure activities and doesn't offer support during her school stress.

Nemotron-3-Nano (verdict: **B**, 45.7 s)

[Accuracy] Summary A says she "wants to break up"; however, it was *he* who suggested breaking up, not her. She expressed frustration and felt underappreciated, but never stated she wants to end the relationship. Summary A therefore introduces a claim not directly supported by the post. [. . .]
Summary B stays strictly within what is stated: frustration, lack of effort in household duties, feeling unappreciated, his focus on leisure. Accuracy is the primary criterion; B is more faithful to the post.

Gemma-4-31B (verdict: **B**, 72.2 s)

[Accuracy] Summary A contains a significant factual error. It states that OP "wants to break up," whereas the text specifies that it was the *boyfriend* who said "we should just break up" during the argument. Summary B is accurate and faithfully reflects the content of the post. [. . .]
Summary A's inaccuracy regarding who suggested the break-up makes it unreliable; B is the clear winner on accuracy.

DeepSeek-R1-Distill-Qwen-32B (verdict: **A**, 58.6 s)

[Accuracy] Summary A mentions that she wants to break up—a specific, emotionally significant detail from the post. Although it was he who voiced the suggestion, the context implies she may be considering it given her frustration; this is arguably a reasonable inference rather than a fabrication. [. . .]
[Helpfulness] Including the break-up detail gives a clearer picture of the severity of her feelings, making Summary A more useful for understanding how serious the situation is.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state contributions being the robust game formulation of alignment with preference ambiguity, the practical algorithm, and experimental evidence, all of which are delivered at the level stated in Section 6.1–6.3.

   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: Key limitations are acknowledged throughout: (1) the hard-constrained robust problem is NP-hard in general, motivating the proxy formulation (Section 4); (2) Tight proxy-to-hard equivalence requires Assumption 3, which depend on algorithmic choices. (3) The proxy introduces a gap of $O\big(\frac{L_\delta}{T} + \delta\big)$, plus the term $\rho\gamma_{\min}$ unless Assumption 3 holds, (4) The LLM experiments use small-scale BT judge instead of production-scale RLHF, as proof of concepts.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All assumptions (Assumptions 1–3) are numbered and explicitly stated. Theorems 1–3 reference the required assumptions in their statements. Full proofs appear in the appendix (Sections B), with all supporting lemmas in Appendix D.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The tabular game construction is fully specified in Section 6. The IMDb experiment is fully detailed in Appendix E. The LLM experiment specifies models used as player/policy, BT judge training, adversary architecture, and evaluation process in Appendix F, including the sampled four-player step (Appendix F.1).

   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: Code is not released at submission time to preserve anonymity. All datasets used (OpenAI TL;DR, IMDb) are publicly available and cited. We will release anonymized code with the supplemental material and the full implementation upon acceptance.

   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: Section 6 provides key details for the tabular and IMDb continuation experiments. Appendix F provides a full hyperparameter table for the experiments, as well as the LLM training details and evaluation protocol.

   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 exact Clopper–Pearson 95% confidence intervals, following the evaluation protocol of prior NLHF work; for $n = 1000$, the worst-case half-width is approximately 0.032. For the IMDb experiments, win rates are averaged over 8000 prompts and 600 sampled kernels per evaluation radius, so the reported values reflect variability over both prompt instances and random kernel perturbations. For the tabular experiment (Figure 1) we report paired-bootstrap 95% intervals over the 60 prompts.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: GPU type and wall-clock time are not currently reported. The tabular and IMDb experiments run on CPU in minutes. The LLM experiments (both training and inference) used 2xH100 GPUs.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The paper studies foundational methodology for robust preference alignment. No sensitive personal data is collected or released; all datasets are publicly available. The work does not involve human subjects research, surveillance, or harmful applications.

   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: [N/A]

    Justification: The paper proposes a foundational optimization framework for robust preference-based alignment and does not introduce a deployed system, new dataset, user-facing model, or application-specific capability. The work is methodological and evaluated only on public benchmark datasets and automated LLM judges. Therefore, there is no direct societal impact from the paper itself.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release new pre-trained models or scraped datasets. The fine-tuned LoRA adapters are lightweight task-specific modules (summarization) with no elevated misuse risk beyond the base models, which are publicly available under their original licenses.

    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 use only existing public datasets and open model checkpoints, all of which are cited in the paper. The IMDb dataset, TL;DR summarization dataset, DistilBERT, Flan-T5, Qwen, Gemma, DeepSeek-R1-Distill-Qwen, and Nemotron models are credited through their corresponding papers or technical reports. We do not redistribute the original datasets or base model weights, and our experiments are conducted for research purposes under the respective licenses and terms of use of these assets.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: All assets are properly cited: the TL;DR dataset [41], IMDb corpus [23], Qwen2.5-1.5B-Instruct [51], Gemma-2-2B [44], DistilBERT [39], and the three external LLM judges. All models are used under their respective open-source licenses consistent with research purposes.

    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: No crowdsourcing or human subject experiments are conducted. All preference data are drawn from the existing public TL;DR and IMDb datasets; external judge evaluations are performed by automated LLM models, not human annotators.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper involves no human subjects research. All experiments use existing public datasets and automated LLM judges.

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 central experimental components. All models are described with version identifiers and citations in Appendices E and F.

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.