ROBUST NASH LEARNING FROM HUMAN FEEDBACK UNDER PREFERENCE UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Nash learning from human feedback (NLHF) models preference alignment as a two-player zero-sum game under general pairwise preferences directly, bypassing the restrictive latent reward assumption. But prior works typically assume this preference is unique and fixed. Yet in practice, such preference can be uncertain or pluralistic. We thus study robust NLHF, seeking a policy that is robust to all plausible preferences within some ambiguity set through optimizing the worst-case performance under the opponent policy and adversarial preference. However, the resulting double minimization does not generally preserve the convex-concave structure of standard NLHF. In this paper, we show that this difficulty can be removed exactly by lifting the adversary to a joint scenario–response law with conditional KL regularization, from which we restore the convex-concave structure and preserve the original robust value. We first consider the case of finite ambiguity sets, showing our reformulation yields a closed-form joint entropic Mirror-Prox algorithm with an $O(1/T)$ convergence guarantee. We then extend the framework beyond finite scenario families to finite-dimensional compact-convex uncertainty sets with affine preference kernels. Although the natural continuous initialization has infinite KL radius against atomic worst cases, we use a localization argument to recover an $O(\log T/T)$ rate. We further evaluate the performance and robustness of our framework. Our studies thus provide a unified optimization framework for robust alignment under uncertain preference beyond latent rewards.

1 INTRODUCTION

Aligning large language models (LLMs) with human preferences is central to making their outputs useful and appropriate for their intended users. In such alignment problems, pairwise comparisons provide a practical learning signal when desired behavior is difficult to specify through an explicit objective. Reinforcement learning from human feedback (RLHF) uses such comparisons to learn a reward model and then optimizes a policy against that model (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). Direct preference optimization further simplifies this pipeline by expressing the solution of a KL-regularized reward objective through a preference loss (Rafailov et al., 2023). These preference-based learning approaches have shown great empirical gains, yet their performance relies on the central assumption that the preference signal can be induced by a latent scalar reward signal, like the Bradley–Terry model. However, such a reward-based representation does not fully capture and model the complicated human preference. For example, a scalar reward induces an ordering of responses, while general pairwise judgments can contain cyclic comparisons. To bypass this restriction and develop more applicable alignment approaches, Nash learning from human feedback (NLHF) is developed as a framework for optimizing such general preferences (Munos et al., 2024; Ye et al., 2024). NLHF uses a general preference model to directly predict the probability that one response is preferred to another, and models the alignment as a game between two response-generating policies: the first policy maximizes its expected preference against the second, which chooses a competing response distribution. The resulting Nash solution accommodates non-transitive preferences, including cycles that can arise when individually transitive judgments are aggregated (Munos et al., 2024).

However, the preference kernel itself is often uncertain. Limited comparisons leave some response pairs poorly estimated, and policy optimization can exploit errors in the fitted model, producing preference hacking (Gupta et al., 2025). Another source of uncertainty arises when evaluators dis-

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

agree or their relative importance changes across deployment settings (Chakraborty et al., 2024). For example, evaluators may place different weights on different criteria, so the comparison probabilities depend on which evaluators are represented. A policy optimized for one fitted kernel can therefore be sensitive to both estimation error and evaluator composition.

Existing approaches already incorporate such robustness into preference alignment. MaxMin-RLHF optimizes performance across a mixture of reward models (Chakraborty et al., 2024), and group robust preference optimization minimizes the worst-group preference loss (Ramesh et al., 2024). For general pairwise preferences, Gupta et al. (2025) study pessimistic Nash objectives and develop practical approximations for preference optimization. These developments motivate the optimization problem studied in this paper: how to compute a policy with a certified worst-case regularized preference value over a prescribed family of general preference kernels. Our focus is the exact reformulation of this game and algorithms with convergence guarantees for both finite and continuous uncertainty families.

We model this problem as robust Nash learning from human feedback. We use an ambiguity set $\mathcal{U}$ to model and index uncertain preference matrices $A(p)$, where $A_{ij}(p)$ is the probability that response i is preferred to response j under preference scenario p. The leader commits to a policy π, and an adversary then chooses both a scenario p and a competing policy ν to minimize the KL-regularized probability that π is preferred to ν under $A(p)$. The ambiguity set can represent a finite collection of preference models or a continuous family, such as uncertain mixtures of evaluator profiles. This additional adversarial choice changes the optimization geometry of standard NLHF. Even when $A(p)$ is affine and $\mathcal{U}$ is convex, the interaction $\pi^\top A(p)\nu$ couples the scenario p and competing policy ν bilinearly. The inner problem need not be jointly convex, so the convex–concave structure of the nominal game does not directly extend to these separate adversarial variables. We resolve this difficulty by representing the adversary through a joint probability law over scenarios and competing responses. The competing policy becomes a conditional response distribution given the scenario. Regularizing this conditional distribution preserves the competing policy's original KL penalty and leaves the scenario marginal unregularized. The resulting payoff is concave in the leader and convex in the joint adversary, enabling us to design convergent algorithms. Specifically, our contributions are as follows.

An exact joint reformulation. The additional minimization over preference kernels need not preserve the saddle-point structure of standard NLHF. To fix this, for every continuous preference family indexed by a nonempty compact metric space, we propose to lift the separate minimizations to a joint law over the product space with a lifted conditional KL regularizer, and prove that this formulation preserves the worst-case value of every leader policy. The key observation is that each joint law averages feasible scenario–policy payoffs, while an atomic law represents a worst pair. We further establish convexity and weak lower semicontinuity of the conditional KL regularizer, minimax duality, and existence of saddle points. Thus we restore the saddle-point geometry of the robust NLHF, which enables us to utilize extensive learning algorithms for zero-sum two-player games.

Optimization and convergence. We further study algorithm design and convergence under two uncertainty set models. For finite $\mathcal{U}$ with K scenarios within, we derive an exact joint entropic proximal update that couples scenario reweighting with the conditional response updates. Building on composite Mirror-Prox (He et al., 2015), we retain both game regularizers inside the proximal subproblems and prove an $O(1/T)$ saddle-gap bound, with logarithmic dependence on K and $O(Km^2)$ arithmetic per round for dense preference matrices; the gap also bounds the leader's robust policy-value error. We then extend the framework to compact convex uncertainty sets in finite-dimensional spaces with affine preference kernels under exact normalization and score integrals. A continuous initialization has infinite KL divergence against atomic worst cases, preventing a direct finite-radius argument. We thus approximate each worst scenario by a localized distribution while keeping its conditional competing policy fixed. Relative-volume scaling gives logarithmic KL cost, and affinity gives a linear payoff error. Balancing these terms yields an $O(\log T/T)$ saddle-gap bound over the full adversarial domain, including atomic and singular joint laws.

Empirical validations. We evaluate fixed-candidate response selection under finite preference uncertainty on PersonalLLM (Zollo et al., 2025) and continuous evaluator-mixture uncertainty on IMDb (Maas et al., 2011). The PersonalLLM study measures the prescribed worst-case regularized value and numerical optimality gaps. The IMDb study evaluates pairwise preference advantages

over nominal Nash-MD under sampled kernel and evaluator-mixture shifts. Both experiments show enhanced robustness under uncertain preferences, validating our theory.

1.1 Related Work

We discuss the closest connections below and provide an extensive review in Appendix A.

Nash alignment with general preferences. Nash learning from human feedback formulates alignment as a zero-sum two-player game between response policies, accommodating preferences that a scalar reward ordering cannot represent (Munos et al., 2024; Swamy et al., 2024). It is further studied under different settings including online preference optimization and self-play setting, with convergent algorithms and statistical learning guarantees (Calandriello et al., 2024; Ye et al., 2024; Zhang et al., 2024). But all these works align to a fixed preference, whereas we introduce uncertainty about the comparison kernel itself, and another player who selects the preference additionally. The adversary jointly selects a preference scenario and a competing response policy, making the inner problem potentially nonconvex, even when the kernel depends affinely on the scenario.

Preference uncertainty and pessimism. Most existing pessimistic alignment methods account for preference uncertainty under BT reward model assumptions, e.g., through adversarial reward optimization, reward-model distillation, suitable policy regularization, or preference uncertainty set centered at a Bradley-Terry model (Liu et al., 2024; Fisch et al., 2024; Huang et al., 2025; Li et al., 2026). The closest objective-level connection is Gupta et al. (2025), who study pessimistic Nash objectives with an adversarial preference model, analyze the effects of limited coverage, and develop restricted-opponent formulations and practical optimization procedures. However, their formulation optimizes a surrogate upper bound rather than the original pessimistic alignment objective exactly. Our contribution is to establish an exact convex–concave reformulation of the robust game, which supports explicit proximal updates and optimization certificates for uncertainty families.

2 Preliminaries

Setup and pairwise preference feedback. Let $\mathcal{X}$ denote the prompt space and let $\mathcal{D}_0$ be a distribution over prompts. For each prompt $x \in \mathcal{X}$, we consider a finite set of candidate responses $\mathcal{Y}(x) = \{y_1, \ldots, y_m\}$. A policy $\boldsymbol{\pi}$ assigns a conditional distribution $\boldsymbol{\pi}(\cdot \mid x) \in \Delta(\mathcal{Y}(x))$; let $\Pi := \{\boldsymbol{\pi} : \boldsymbol{\pi}(\cdot \mid x) \in \Delta(\mathcal{Y}(x)), x \in \mathcal{X}\}$ denote the full policy class. We fix a reference policy $\boldsymbol{r}(\cdot \mid x)$ with full support, representing the pre-alignment model from SFT, to which policy updates will be regularized.

Human feedback is observed through pairwise comparisons rather than absolute scores. Given a prompt x and two responses $y, y' \in \mathcal{Y}(x)$, let $P_x(y \succ y') := \Pr(y \text{ is preferred to } y' \mid x)$ denote their pairwise preference probability. Equivalently, a preference query produces a binary observation $z \sim \mathrm{Ber}(P_x(y \succ y'))$, where $z = 1$ indicates that y is preferred to y' and $z = 0$ indicates the reverse outcome. Such pairwise preference models are standard in general-preference alignment (Munos et al., 2024; Ye et al., 2024).

We consider general pairwise preferences satisfying the following reciprocity condition, as commonly assumed in game-theoretic preference alignment (Munos et al., 2024; Ye et al., 2024).

Assumption 1 (Reciprocal preferences)**.** *For any fixed prompt x, define $P_x(y \succ y') := P(y \succ y' \mid x)$. We assume that the preference probabilities are reciprocal, i.e., for every prompt x and every pair $y, y' \in \mathcal{Y}(x)$, including $y = y'$, assume $P_x(y \succ y') + P_x(y' \succ y) = 1$.*

We allow any reciprocal pairwise preference kernel. A common special case is the Bradley–Terry model (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022), under which pairwise preferences are induced by a scalar reward function $R : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$: $P_x^{\mathrm{BT}}(y \succ y') = \frac{\exp(R(x,y))}{\exp(R(x,y)) + \exp(R(x,y'))}$. Reward-induced models of this form constitute a restricted subclass of general pairwise preference models (Munos et al., 2024; Ye et al., 2024). In particular, comparison probabilities need not in general admit any representation through a single scalar reward. They may, for example, exhibit a preference cycle $P_x(y_1 \succ y_2) > \frac{1}{2}$, $P_x(y_2 \succ y_3) > \frac{1}{2}$, and $P_x(y_3 \succ y_1) > \frac{1}{2}$. General pairwise preferences therefore provide the natural representation for game-theoretic alignment when the latent-reward assumption is not imposed.

Nash preference alignment. A pairwise preference model naturally induces a preference relation between response-generating policies. Given two policies $\boldsymbol{\pi}, \boldsymbol{\nu} \in \Pi$, define the policy-level preference probability under some preference kernel P as

$$P(\boldsymbol{\pi} \succ \boldsymbol{\nu}) := \mathbb{E}_{x\sim\mathcal{D}_0}\mathbb{E}_{y\sim\boldsymbol{\pi}(\cdot|x),\, y'\sim\boldsymbol{\nu}(\cdot|x)}\big[P_x(y \succ y')\big]. \tag{1}$$

Thus, $P(\boldsymbol{\pi} \succ \boldsymbol{\nu})$ is the expected probability that a response generated by $\boldsymbol{\pi}$ is preferred to an independently generated response from $\boldsymbol{\nu}$ under the same prompt. Following Nash learning from human feedback (Munos et al., 2024; Ye et al., 2024), for $\tau > 0$, alignment under general preferences is formulated as a regularized two-player zero-sum game $\max_{\boldsymbol{\pi}\in\Pi}\min_{\boldsymbol{\nu}\in\Pi} J(\boldsymbol{\pi}, \boldsymbol{\nu}; P)$, w.r.t. the KL-regularized pay-off function:

$$J(\boldsymbol{\pi}, \boldsymbol{\nu}; P) := P(\boldsymbol{\pi} \succ \boldsymbol{\nu}) - \tau\mathbb{E}_{x\sim D_0}\big[D_{\mathrm{KL}}\big(\boldsymbol{\pi}(\cdot \mid x)\|\boldsymbol{r}(\cdot \mid x)\big)\big] + \tau\mathbb{E}_{x\sim D_0}\big[D_{\mathrm{KL}}\big(\boldsymbol{\nu}(\cdot \mid x)\|\boldsymbol{r}(\cdot \mid x)\big)\big].$$

In the game, $\boldsymbol{\pi}$ is the max-player and $\boldsymbol{\nu}$ is the min-player. The max-player seeks a policy whose responses are consistently preferred to those of a competitor, while the min-player selects a competing response distribution that makes this performance as small as possible. The nominal general-preference alignment problem is therefore formulated as the regularized minimax game seeking a Nash equilibrium, or equivalently a saddle pair $(\boldsymbol{\pi}^\star, \boldsymbol{\nu}^\star)$ satisfying

$$J(\boldsymbol{\pi}, \boldsymbol{\nu}^\star; P) \le J(\boldsymbol{\pi}^\star, \boldsymbol{\nu}^\star; P) \le J(\boldsymbol{\pi}^\star, \boldsymbol{\nu}; P) \qquad \text{for every feasible } \boldsymbol{\pi}, \boldsymbol{\nu}. \tag{2}$$

In the following analysis, we consider pointwise robust optimization across prompts: each prompt defines an independent full-simplex game, and the adversary may choose its preference scenario and competing response distribution separately at each prompt. We therefore fix $x \in \mathcal{X}$ and suppress its dependence in the notation. The preference model at this prompt is represented by $A \in [0,1]^{m\times m}$ with $A_{ij} := P_x(y_i \succ y_j)$. The conditional policy-level preference reduces to $\mathbb{E}_{y\sim\pi,\, y'\sim\nu}\big[P_x(y \succ y')\big] = \pi^\top A\nu$. Accordingly, we write the fixed-prompt NLHF payoff as

$$J(\pi, \nu; A) := \pi^\top A\nu - \tau D_{\mathrm{KL}}(\pi\|r) + \tau D_{\mathrm{KL}}(\nu\|r), \qquad \pi, \nu \in \Delta_m. \tag{3}$$

Reciprocity implies $J(\pi, \nu; A) + J(\nu, \pi; A) = 1$. Moreover, for each fixed prompt, the full-simplex game is strictly concave-convex, and therefore max-min duality holds and it has a unique saddle point. The reciprocity identity and the common reference imply that this saddle is symmetric, $(\pi_A^\star, \pi_A^\star)$, with value $J(\pi_A^\star, \pi_A^\star; A) = \frac{1}{2}$. (Munos et al., 2024). For the prompt-averaged full-policy formulation, uniqueness is understood up to $\mathcal{D}_0$-almost-sure equality.

3 Robust Nash Alignment: An Exact Joint Formulation

Optimizing a fitted preference kernel can exploit its estimation errors, favoring responses with spuriously high predicted preferences. This phenomenon, known as preference hacking, is particularly relevant when comparison data provide limited coverage. KL regularization controls policy drift but does not explicitly model uncertainty in preference predictions (Gupta et al., 2025). In other cases, the preference itself can be uncertain or diverse (Chakraborty et al., 2024). Motivated by pessimistic preference optimization, we therefore study the worst-case regularized payoff over a prescribed family of preference kernels rather than a single estimate.

At the fixed prompt, both policies range over Δ_m, and r is the full-support reference. Recall the fixed-prompt payoff $J(\pi, \nu; A)$ from equation 3. We let $\mathcal{U}$ be a nonempty compact metric space which represents the potential or plausible preference under uncertainty, and let $A : \mathcal{U} \to [0,1]^{m\times m}$ be a continuous family of reciprocal preference matrices. This set can cover different uncertainties. For example, a finite family ($|\mathcal{U}| < \infty$) can represent several fitted preference models or evaluator profiles; A continuous set can represent uncertain evaluator weights over different criteria: $A(p) := \sum_{k=1}^K p_k A^{(k)}$ with p in a compact subset of Δ_K. Extending the NLHF formulation, our robust NLHF formulation seeks a policy that maximizes the worst-case payoff under both preference and ν:

$$g(\pi) := \min_{p\in\mathcal{U}, \nu\in\Delta_m} J(\pi, \nu; A(p)), \qquad V^\star := \max_{\pi\in\Delta_m} g(\pi). \tag{4}$$

Note that for every deployment scenario $p^{\mathrm{dep}} \in \mathcal{U}$ with kernel $A(p^{\mathrm{dep}})$ and every competitor ν, $J(\pi, \nu; A(p^{\mathrm{dep}})) \ge g(\pi)$, so $V^\star$ is the largest regularized floor that a policy can guarantee uniformly.

The optimization challenge. Although equation 4 only adds an additional minimization layer to standard NLHF, it fundamentally changes the problem structure. For each fixed (p, ν), the payoff is concave in π, so g is concave. The key difficulty lies in the adversary's parameterization: even when $\mathcal{U}$ is convex and $A(p)$ is affine, $\pi^\top A(p)\nu$ couples p and ν bilinearly and need not be jointly convex in these variables. Eliminating the opponent by the Gibbs variational identity gives

$$g(\pi) = -\tau D_{\mathrm{KL}}(\pi\|r) + \min_{p\in\mathcal{U}} \left\{ -\tau \log \sum_{j=1}^{m} r_j \exp\left(-\frac{[A(p)^\top \pi]_j}{\tau}\right) \right\}. \tag{5}$$

For affine A, the expression inside braces is concave in p. This exact elimination therefore leaves a concave minimization problem. The separate adversarial variables need not satisfy max–min equality; Remark 2 gives an explicit strict gap. Consequently, the assumptions underlying nominal saddle-point analyses do not apply directly.

In prior work, Gupta et al. (2025) replace the objective by a variational upper-bound surrogate and optimize the resulting policy–preference game using gradient ascent–descent. Their P3O algorithm uses an opponent formed from an exponential moving average of past policies and regularizes the preference model toward its fitted estimate. These approximations enable tractable stochastic updates, but maximizing the surrogate need not recover the original robust optimum. We instead construct a joint adversarial variable that preserves $g(\pi)$ exactly and exposes a convex–concave saddle problem with convergence guarantees.

3.1 Joint laws and conditional regularization

The difficulty discussed above is that the adversary jointly chooses the scenario p and the competing policy ν, and treating them as separate optimization variables does not in general yield a convex inner problem. We thus instead propose to combine the two together, and to encode their joint choice by a joint probability law. Let $\mathcal{Q} := \mathcal{P}(\mathcal{U} \times [m])$ be the Borel probability laws over scenarios and competing responses. Every $Q \in \mathcal{Q}$ admits a scenario marginal α_Q and a conditional response distribution $\nu_Q(p) \in \Delta_m$, with $Q(dp, j) = \alpha_Q(dp)\nu_{Q,j}(p)$. Define

$$s(Q) := \sum_{j=1}^{m} \int_{\mathcal{U}} A(p)e_j \, Q(dp, j), \qquad \Psi(Q) := \int_{\mathcal{U}} D_{\mathrm{KL}}(\nu_Q(p)\|r)\, d\alpha_Q(p), \tag{6}$$

where e_j is the jth coordinate vector. We then define the lifted payoff as

$$L(\pi, Q) := \pi^\top s(Q) - \tau D_{\mathrm{KL}}(\pi\|r) + \tau\Psi(Q). \tag{7}$$

The vector $s(Q)$ gives the preference score of every leader response against the joint adversary. The conditional KL term charges each scenario for the competing policy used in that scenario.

Remark 1 (Why the regularizer must be conditional)**.** *The KL chain rule gives $\Psi(Q) = D_{\mathrm{KL}}(Q\|\alpha_Q \otimes r)$. Consequently, the scenario marginal is unpenalized and may concentrate on a worst scenario. Replacing this term by KL to a fixed product law $\beta \otimes r$ would add $D_{\mathrm{KL}}(\alpha_Q\|\beta)$, changing the robust objective. With a diffuse β, it would also assign infinite cost to atomic scenario choices. The moving marginal in $\alpha_Q \otimes r$ is what preserves the original opponent penalty.*

Note that this formulation covers finite and continuous scenario spaces uniformly. When $|\mathcal{U}| = K$ is finite, a joint law Q is identified with $q \in \Delta_{Km}$ by $q_{kj} = Q(\{k\} \times \{j\})$; the finite variable used in the next section is therefore simply the coordinate representation of the same joint object.

Theorem 3.1 (Exact joint reformulation)**.** *Under the compact-metric, continuous-kernel setup above, the fixed-policy values agree: $g(\pi) = \min_{Q\in\mathcal{Q}} L(\pi, Q)$ for every $\pi \in \Delta_m$. Moreover, the min-max duality holds:*

$$V^\star = \max_{\pi\in\Delta_m} \min_{Q\in\mathcal{Q}} L(\pi, Q) = \min_{Q\in\mathcal{Q}} \max_{\pi\in\Delta_m} L(\pi, Q). \tag{8}$$

All extrema are attained, and any pair of outer optimizers forms a saddle point of L.

For fixed π, $L(\pi, Q) = \int_{\mathcal{U}} J(\pi, \nu_Q(p); A(p))\, d\alpha_Q(p) \geq g(\pi)$, while an atomic worst-pair law $\delta_{p^\star} \otimes \nu^\star$ attains equality. Moreover, $\Psi(Q) = D_{\mathrm{KL}}(Q\|\alpha_Q \otimes r)$ is convex and weakly lower semicontinuous; together with the affine continuity of $s(Q)$ and compactness of $\mathcal{Q}$, this yields the minimax equality.

Remark 2 (A strict minimax gap in the original variables)**.** *We highlight that our lifting and the conditional regularizer are key to restoring the saddle point structure. Consider an example: let* $r = (1/2, 1/2)$*,* $0 < \delta \le 1/2$*, and* $A(\theta) = \begin{pmatrix} 1/2 & 1/2+\theta \\ 1/2-\theta & 1/2 \end{pmatrix}$ *for* $\theta \in [-\delta, \delta]$*. Write* $\pi = (u, 1-u)$ *and* $\ell(u) = D_{\mathrm{KL}}((u, 1-u)\|r)$*. Equation 5 gives* $g(\pi) = 1/2 - \delta|u - 1/2| - \tau\ell(u) - \tau \log\cosh(\delta/(2\tau))$*. Hence the robust leader is* r *and* $V^\star = 1/2 - \tau \log\cosh(\delta/(2\tau)) < 1/2$*. In contrast,* $\min_{\theta,\nu} \max_\pi J(\pi, \nu; A(\theta)) = 1/2$ *because each fixed-scenario nominal game has value* $1/2$*. The joint-law minimax equality therefore repairs a genuine failure of the exchange in the original variables. An atomic law suffices for a fixed leader's worst reply, whereas a saddle adversary can require mixing scenarios with different conditional opponents.*

Certificates for the robust policy. Our saddle formulation further provides both a policy and an upper bound on the best achievable robust value. Define $h(Q) := \max_{\pi' \in \Delta_m} L(\pi', Q)$. A second Gibbs calculation gives $h(Q) = \tau\Psi(Q) + \tau \log \sum_{i=1}^m r_i e^{s_i(Q)/\tau}$. Note that for every feasible pair, $g(\pi) \le V^\star \le h(Q)$, thus

$$0 \le V^\star - g(\pi) \le \mathrm{Gap}(\pi, Q), \text{ where } \mathrm{Gap}(\pi, Q) := h(Q) - g(\pi). \tag{9}$$

Thus a gap of at most ε certifies that the leader's worst-case regularized preference value is within ε of optimal. The certificate also controls the leader policy itself. At a saddle $(\pi^\star, Q^\star)$, the unique leader best response satisfies $\pi_i^\star \propto r_i e^{s_i(Q^\star)/\tau}$ and has full support. The Gibbs identity gives $L(\pi^\star, Q^\star) - L(\pi, Q^\star) = \tau D_{\mathrm{KL}}(\pi\|\pi^\star)$. Since $g(\pi) \le L(\pi, Q^\star)$, this implies $\tau D_{\mathrm{KL}}(\pi\|\pi^\star) \le V^\star - g(\pi) \le \mathrm{Gap}(\pi, Q)$. In particular, the robust leader is unique, and Pinsker's inequality yields $\|\pi - \pi^\star\|_1 \le \sqrt{2\,\mathrm{Gap}(\pi, Q)/\tau}$. Therefore, robust NLHF reduces to finding a saddle point of $L(\pi, Q)$.

4 Exact Joint Mirror-Prox for Finite Scenarios

In this section, we investigate convergent algorithms for the robust NLHF. We start with a simple case of finite uncertainty set. Such sets can model the preference from different groups, and our robust objective provides an optimized lower bound for alignment with them. A similar problem is studied in (Chakraborty et al., 2024) under the latent reward assumption, thus our studies here generalize it to NLHF formulation.

Let $\mathcal{U} = [K]$ with matrices $A^1, \ldots, A^K \in [0,1]^{m \times m}$. The joint adversary is represented by $q \in \Delta_{Km}$, with scenario masses $\alpha_k = \sum_j q_{kj}$ and conditional policies $q_{kj} = \alpha_k \nu_{kj}$. Then $\Psi(q) = \sum_{k,j} q_{kj} \log(q_{kj}/(\alpha_k r_j))$, with zero-mass rows contributing zero. The interaction in $L(\pi, q)$ is determined by the leader scores $s_i(q) = \sum_{k,j} q_{kj} A^k_{ij}$ and adversarial costs $c_{kj}(\pi) = \sum_i \pi_i A^k_{ij}$.

4.1 A closed-form joint proximal update

We adopt the composite Mirror-Prox method (He et al., 2015). Rather than differentiating every term in L, we evaluate the bilinear interaction and retain both KL regularizers inside the proximal subproblems. This exploits their tractable entropic structure and avoids requiring globally Lipschitz entropy gradients near the simplex boundary. For an old center $z^\circ = (\pi^\circ, q^\circ)$ and evaluated scores and costs (s, c), define $\mathrm{JProx}_\eta(z^\circ; s, c) := (\pi^+, q^+)$ by

$$\begin{aligned} \pi^+ &= \arg\max_{\pi \in \Delta_m} \left\{ \eta s^\top \pi - \eta\tau D_{\mathrm{KL}}(\pi\|r) - D_{\mathrm{KL}}(\pi\|\pi^\circ) \right\}, \\ q^+ &= \arg\min_{q \in \Delta_{Km}} \left\{ \eta\langle c, q\rangle + \eta\tau\Psi(q) + D_{\mathrm{KL}}(q\|q^\circ) \right\}. \end{aligned} \tag{10}$$

Here $\langle c, q\rangle = \sum_{k,j} c_{kj} q_{kj}$. The terms multiplied by $\eta\tau$ are the original game regularizers; the centered KL terms define the optimization step and vanish from the target game. This distinction permits regularization of the scenario updates without penalizing the scenario marginal in the robust objective. We further prove that our update has the following closed-form rules.

Proposition 4.1 (Closed-form joint entropic prox)**.** *Let* $q^\circ_{kj} = \alpha^\circ_k \nu^\circ_{kj}$ *have strictly positive entries, and let* π° *be strictly positive. For finite scores and costs and any finite* $\eta > 0$*, let* $a = 1 + \eta\tau$*,* $b_{kj} := (\nu^\circ_{kj})^{1/a} r_j^{\eta\tau/a} e^{-\eta c_{kj}/a}$*, and* $Z_k := \sum_j b_{kj}$*. Then the unique solution of equation 10 is*

$$\pi_i^+ \propto (\pi_i^\circ)^{1/a} r_i^{\eta\tau/a} e^{\eta s_i/a}, \qquad \nu^+_{kj} = \frac{b_{kj}}{Z_k}, \qquad \alpha^+_k = \frac{\alpha^\circ_k Z_k^a}{\sum_{\ell=1}^K \alpha^\circ_\ell Z_\ell^a}. \tag{11}$$

Here π^+ is normalized over i, and $q^+_{kj} = \alpha^+_k \nu^+_{kj}$.

Optimizing the conditional response in scenario k produces the value $-a \log Z_k$. The remaining entropic update over scenario masses therefore gives $\alpha^+_k \propto \alpha^\circ_k Z^a_k$. Thus scenario reweighting incorporates the optimized conditional opponent.

4.2 Predictor–corrector iteration: algorithm and convergence.

Using the closed-form map in Proposition 4.1, Algorithm 1 applies it twice per iteration: the predictor uses the current scores and costs, and the correction reevaluates them at the predictor while retaining the original proximal center. Initialize $\pi_1 = r$ and $q_{1,kj} = r_j/K$. At round t, set $\widetilde{z}_t = \mathrm{JProx}_\eta(z_t; s(q_t), c(\pi_t))$ and $z_{t+1} = \mathrm{JProx}_\eta(z_t; s(\widetilde{q}_t), c(\widetilde{\pi}_t))$. Both steps use the same proximal center z_t. Return $(\bar{\pi}_T, \bar{q}_T) = T^{-1} \sum_{t=1}^T \widetilde{z}_t$.

Theorem 4.2 (Finite-scenario convergence)**.** *Let $r_{\min} := \min_j r_j$ and $R := \log(1/r_{\min})$. For unit-range preference matrices and exact updates, Algorithm 1 with a fixed $0 < \eta \le 1$ satisfies*

$$0 \le V^\star - g(\bar{\pi}_T) \le \mathrm{Gap}(\bar{\pi}_T, \bar{q}_T) \le \frac{\log K + 2R}{\eta T} \qquad (T \ge 1). \tag{12}$$

In particular, $\eta = 1/2$ gives the bound $2(\log K + 2R)/T$.

The bilinear interaction is 1-Lipschitz in the product geometry used by Mirror-Prox, while both entropy terms remain inside the proximal step. Consequently, for every comparison pair $(\pi', q') \in \Delta_m \times \Delta_{Km}$, we have $L(\pi', \bar{q}_T) - L(\bar{\pi}_T, q') \le \frac{D_{\mathrm{KL}}(\pi' \| r) + D_{\mathrm{KL}}(q' \| q_1)}{\eta T}$. The initial KL radius is at most $\log K + 2R$; maximizing over π' and minimizing over q' yields Theorem 4.2.

Remark 3 (Complexity)**.** *With $\eta = 1/2$, any integer $T \ge 1$ satisfying $T \ge \lceil 2(\log K + 2R)/\varepsilon \rceil$ suffices for an ε saddle gap. Each iteration costs $O(Km^2)$ operations for dense matrices, including the two interaction evaluations, and $O(Km)$ additional operations for the proximal updates. The scenario dependence is logarithmic in the iteration bound and linear in each iteration's arithmetic cost. For uniform r, the numerator is $\log K + 2 \log m$.*

Proximal KL treatment and extragradient updates already appear in nominal NLHF and entropy-regularized games (Cen et al., 2021; Tiapkin et al., 2025). The additional structure here is the conditional penalty on a joint scenario–response law. In particular, $\Psi(\alpha \otimes r) = 0$ for every scenario distribution α, so this penalty supplies no curvature along scenario-marginal directions. The strong convexity used for linear convergence in fully regularized games therefore does not carry over directly. Our $O(1/T)$ guarantee follows from composite Mirror-Prox after deriving the exact joint prox and verifying the required geometry, thereby solving the exact robust NLHF problem.

5 Continuous Uncertainty and Full-Domain Convergence

We next consider the continuous uncertainty set case. Let $\mathcal{U} \subset \mathbb{R}^D$ be nonempty, compact, and convex, and let $A : \mathcal{U} \to [0,1]^{m \times m}$ be affine and unit-range. If $\mathcal{U}$ is a polytope with explicitly enumerated vertices, affine scenario minimization reduces to those vertices and the finite algorithm applies; here we treat general compact-convex sets, including those with infinitely many extreme points. Let $d_{\mathcal{U}} = \dim \mathrm{aff}(\mathcal{U})$ (e.g., $d_{\mathcal{U}} \le K - 1$ for $\mathcal{U} \subseteq \Delta_K$). For $d_{\mathcal{U}} \ge 1$, let $\lambda_{\mathcal{U}}$ be normalized $d_{\mathcal{U}}$-dimensional volume on $\mathcal{U}$ within its affine hull. For $d_{\mathcal{U}} = 0$, use the Dirac law on the singleton. Theorem 3.1 continues to apply, with the full domain $\mathcal{Q} = \mathcal{P}(\mathcal{U} \times [m])$.

5.1 The measure-valued update and its access model

For the continuous case, we can similarly design an algorithm following the proximal methods. Initialize $\pi_1 = r$ and $Q_1 = \lambda_{\mathcal{U}} \otimes r$. For a center $z^\circ = (\pi^\circ, Q^\circ)$ with $Q^\circ(dp, j) = \alpha^\circ(dp)\nu^\circ_j(p)$ and evaluated scores and costs (s, c), similarly, the pair is updated by

$$\pi^+ = \arg\max_{\pi \in \Delta_m} \left\{ \eta s^\top \pi - \eta\tau D_{\mathrm{KL}}(\pi \| r) - D_{\mathrm{KL}}(\pi \| \pi^\circ) \right\}, \tag{13}$$

$$Q^+ = \arg\min_{Q \in \mathcal{Q}} \left\{ \eta \langle c, Q \rangle + \eta\tau \Psi(Q) + D_{\mathrm{KL}}(Q \| Q^\circ) \right\}, \tag{14}$$

where $\langle c, Q\rangle := \sum_{j=1}^{m} \int_{\mathcal{U}} c_j(p)\, Q(dp, j)$ and $c(p) := A(p)^\top \pi$. We further prove that the adversarial update Q^+ in equation 14 admits a closed-form conditional response update pointwise in p, presented in Lemma C.2. We thus design our concrete Exact Joint Entropic Mirror-Prox algorithm following these updates (EJMP), presented in Algorithm 2. In our algorithm, we assume access to the normalized reference law, exact pointwise operations on the kernel and generated laws, the initial score, and the integrals in the update rules.

5.2 Localizing a worst reply without changing its KL penalty

Failure of the finite-radius bound. Although algorithms under both settings are prox-based, the convergence analysis cannot be applied. The finite-scenario convergence proof relies on a uniform bound on the initial KL radius. Indeed, with $q_{1,kj} = r_j/K$, every scenario response atom has positive mass, so for any comparison pair (π', Y), $D_{\mathrm{KL}}(\pi'\|r) \le R$ and $D_{\mathrm{KL}}(Y\|q_1) \le \log K + R$, which gives the finite radius $\log K + 2R$ in Theorem 4.2. For $d_{\mathcal{U}} \ge 1$, however, the reference scenario law is diffuse. For a fixed leader, Theorem 3.1 gives a worst pair $(p^\star, \nu^\star)$, whose joint law is $\delta_{p^\star} \otimes \nu^\star$. Since $\lambda_{\mathcal{U}}(\{p^\star\}) = 0$ whereas $\delta_{p^\star}(\{p^\star\}) = 1$, this atomic law assigns positive mass to a set that $Q_1 = \lambda_{\mathcal{U}} \otimes r$ assigns zero mass. Hence $D_{\mathrm{KL}}(\delta_{p^\star} \otimes \nu^\star\|Q_1) = \infty$, making the finite-scenario initial-radius bound vacuous. Restricting the comparison to a fixed density cap would change the gap being certified. We thus resolve this obstruction by localizing the comparison law and approximating a worst reply in the proof.

A finite-entropy comparison law. To obtain a finite comparison radius, fix the output $\bar\pi_T$ and let $(p^\star, \nu^\star)$ be a worst pair for it. For $0 < \sigma \le 1$, we contract $\mathcal{U}$ toward $p^\star$ by the map $\Gamma_\sigma(p) := (1-\sigma)p^\star + \sigma p$, and define $\alpha_\sigma := (\Gamma_\sigma)_\# \lambda_{\mathcal{U}}$ and $Y_\sigma := \alpha_\sigma \otimes \nu^\star$. Here $(\Gamma_\sigma)_\#$ denotes pushforward of a probability law. Convexity ensures feasibility, including when $p^\star$ lies on the boundary. Unlike the atomic law $\delta_{p^\star}$, α_σ remains absolutely continuous with respect to $\lambda_{\mathcal{U}}$, while its support is a σ-contracted copy of $\mathcal{U}$ around $p^\star$. The contraction scales $d_{\mathcal{U}}$-dimensional relative volume by $\sigma^{d_{\mathcal{U}}}$, so $d\alpha_\sigma/d\lambda_{\mathcal{U}} = \sigma^{-d_{\mathcal{U}}} \mathbf{1}_{\Gamma_\sigma(\mathcal{U})}$. Thus $D_{\mathrm{KL}}(\alpha_\sigma\|\lambda_{\mathcal{U}}) = d_{\mathcal{U}} \log(1/\sigma)$, and the KL chain rule gives

$$D_{\mathrm{KL}}(Y_\sigma\|Q_1) = d_{\mathcal{U}} \log(1/\sigma) + D_{\mathrm{KL}}(\nu^\star\|r). \tag{15}$$

Since both $D_{\mathrm{KL}}(\pi'\|r)$ and $D_{\mathrm{KL}}(\nu^\star\|r)$ are at most R, the total initial comparison radius is therefore at most $2R + d_{\mathcal{U}} \log(1/\sigma)$. The response distribution remains $\nu^\star$ at every scenario, so $\Psi(Y_\sigma) = D_{\mathrm{KL}}(\nu^\star\|r)$. Both objective regularizers are therefore unchanged by localization. By affinity, $L(\bar\pi_T, Y_\sigma) - g(\bar\pi_T) = \sigma \int_{\mathcal{U}} \bar\pi_T^\top [A(p) - A(p^\star)] \nu^\star\, d\lambda_{\mathcal{U}}(p)$. Worst-pair optimality and the unit payoff range then give

$$0 \le L(\bar\pi_T, Y_\sigma) - g(\bar\pi_T) \le \sigma. \tag{16}$$

Applying the Mirror-Prox comparison bound to Y_σ gives, for every $\pi' \in \Delta_m$, we have that $L(\pi', \bar Q_T) - L(\bar\pi_T, Y_\sigma) \le \frac{2R + d_{\mathcal{U}} \log(1/\sigma)}{\eta T}$. Maximizing over π' and adding the localization error in Eq. (16) yields the full-domain gap bound below. The localized law Y_σ is only a finite-KL comparison law; the robust value $g(\bar\pi_T)$ still minimizes over the full adversarial domain.

Theorem 5.1. *Under the compact-convex, affine, unit-range, and exact-access assumptions above, Algorithm 2 with fixed $0 < \eta \le 1$ satisfies, for $d_{\mathcal{U}} \ge 1$, every integer $T \ge 1$, and every $0 < \sigma \le 1$,*

$$0 \le \mathrm{Gap}(\bar\pi_T, \bar Q_T) \le \frac{2R + d_{\mathcal{U}} \log(1/\sigma)}{\eta T} + \sigma. \tag{17}$$

In particular, $\eta = 1/2$ and $\sigma = 1/T$ give $\mathrm{Gap}(\bar\pi_T, \bar Q_T) \le (4R + 2d_{\mathcal{U}} \log T + 1)/T = O(\log(T)/T)$. For $d_{\mathcal{U}} = 0$, the bound is $2R/(\eta T)$.

For $0 < \epsilon \le 1$, choosing $\eta = 1/2$ and using $\sigma = \epsilon/2$ in the analysis shows that $T \ge \lceil (8R + 4d_{\mathcal{U}} \log(2/\epsilon))/\epsilon \rceil$ iterations suffice. Thus our approach overcomes the infinite bound issue, and ensures finite convergence rate.

6 Experiments

We further numerically evaluate our framework under both finite and continuous preference uncertainty. We defer the finite-scenario study to Appendix D.1, and mainly present the experiments on IMDb continuation task with continuous uncertainty over heterogeneous evaluator mixtures here.

Table 1: IMDb win fractions against matched-reference Nash-MD, using the SFT-derived reference and $T = 2048$. Columns within each block indicate ρ_{eval}. Values above 0.5 indicate preference advantages on a majority of sampled prompt–kernel pairs.

	KL perturbations					Annotator-mixture shifts				
ρ_{train}	.05	.10	.20	.30	.50	.05	.10	.20	.30	.50
.005	.999	.995	.982	.968	.938	.999	.992	.968	.938	.869
.010	.999	.995	.983	.969	.940	.999	.993	.969	.938	.866
.020	.999	.995	.984	.970	.942	.999	.993	.969	.937	.863
.050	.998	.994	.983	.971	.943	.997	.991	.969	.936	.859
.100	.992	.988	.977	.965	.939	.990	.984	.961	.929	.853
.150	.985	.981	.971	.958	.932	.982	.975	.954	.922	.850

Task and setup. We study response selection under changing evaluator preferences using $N = 4{,}000$ prompts from the negative-label portion of IMDb (Maas et al., 2011). We follow the *best-of-m* protocol to reduce the response space, where we generated $m = 8$ candidate continuations per prompt by FLAN-T5-Large (Chung et al., 2024). The preference is based on three heuristic profiles, favoring detailed analysis, conciseness, and balanced discussion, inducing the preference kernels A_x^1, A_x^2, A_x^3. We then model the continuous preference uncertainty through re-weighting: $A_x(w) := \sum_{g=1}^{3} w_g A_x^g$, with $w \in \mathcal{U}_\rho := \{w \in \Delta_3 : D_{\mathrm{KL}}(w\|w_0) \le \rho\}$ and nominal weights $w_0 = (0.5, 0.3, 0.2)$. This uncertainty set satisfies the assumptions, and our EJMP optimizes a distribution over each prompt's fixed candidates, using weighted quadrature to approximate the continuous integrals, under the training radius ρ_{train}. We use $\tau = 0.01$ and a reference derived from the frozen generator's candidate likelihoods (SFT-derived). The main results use predictor-averaged policies after $T = 2048$ iterations, compared with the nominal Nash-MD baseline using the same reference and regularization.

We consider two stress tests: *KL perturbations* of pairwise logits, controlled by summed Bernoulli–KL divergence, and *annotator-mixture shifts*, obtained by Dirichlet reweighting with rejection-based feasibility checks (Appendix D.2.4). Here ρ_{train} is a KL radius in evaluator-weight space, whereas ρ_{eval} is a summed Bernoulli–KL radius in kernel space; the two are not directly comparable. For each test and evaluation radius, we sampled $M = 600$ deployment/evaluation kernels per prompt from the corresponding stress-test distribution. For a sampled kernel $P_{x,\ell}$, let $W_{x,\ell} := \Pr_{i\sim\pi_x,\, j\sim\pi_{N,x}}[i \succ j \mid P_{x,\ell}] = \pi_x^\top P_{x,\ell}\pi_{N,x}$, i.e., the probability that our policy π_x is preferred over the baseline $\pi_{N,x}$ from NLHF, and we report $\mathrm{WinFrac} := \frac{1}{NM}\sum_{x=1}^{N}\sum_{\ell=1}^{M}\mathbf{1}\{W_{x,\ell} > \frac{1}{2}\}$, the fraction of sampled prompt–kernel pairs for which the robust policy is preferred. We develop our experiment under different training and evaluating radii and report Table 1.

As the results show, EJMP is preferred under most sampled kernels at every tested training–evaluation radius pair. At the largest evaluation radius, win fractions remain at least 0.932 under KL perturbations and 0.850 under annotator-mixture shifts. Averaging equally over the 30 radius pairs in each test yields win fractions of 0.974 and 0.948. These results show that EJMP retains a preference advantage across the sampled evaluator shifts. We defer additional experiment results and comparisons, with the implementation details, control comparisons, iteration-budget analyses, and paired prompt-bootstrap intervals to Appendix D.2.

7 Conclusion

We establish an optimization framework for robust Nash alignment under a prescribed family of uncertain preferences. Representing the adversary by a joint scenario–response law with conditional KL regularization preserves leader policy's robust value and yields a convex–concave saddle problem. The resulting joint proximal map supports an explicit Mirror-Prox algorithm with an $O(1/T)$ gap for finite scenario families. For compact-convex affine uncertainty, localizing a worst scenario while preserving its conditional response policy gives an $O(\log T/T)$ guarantee over the full adversarial domain under exact moment access. These bounds certify the leader's worst-case regularized preference value, and provide a concrete foundation for alignment to uncertain and ambiguous preferences beyond reward-based ones.

AI USE STATEMENT

We used generative AI tools to assist with language editing and proofreading, code debugging, and checking mathematical derivations and experimental workflows. We also used generative AI tools to assist with organizing and summarizing relevant literature.

All AI-assisted content was reviewed and verified by the authors. Mathematical claims and proofs were checked independently against the stated assumptions and derivations; AI-assisted code was executed and validated through numerical checks and controlled experiments; reported results were verified against the underlying experimental outputs; and citations and related-work claims were checked against the original sources. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not collect new human-subject data. Our experiments use existing datasets and model-generated candidate responses to study optimization under uncertain preference models. The proposed framework provides robustness only with respect to a prescribed ambiguity set; it does not by itself guarantee that this set accurately represents the preferences of a deployment population or resolves broader issues of bias, fairness, or preference aggregation. Consequently, care is required when constructing preference scenarios and interpreting the resulting robust policies in real-world applications.

REPRODUCIBILITY STATEMENT

We provide explicit algorithms and optimization assumptions in Sections 4–5, with complete proofs of the theoretical results and auxiliary lemmas in Appendices B and C. Appendix D describes the experimental data construction, preference-kernel construction, optimization settings, evaluation protocols, and additional results for both the finite- and continuous-uncertainty experiments. The reported experiments use fixed candidate sets and specified optimization horizons, references, uncertainty radii, and evaluation procedures to facilitate independent reproduction.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4447–4455. PMLR, 2024.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/2334029`.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 27952–27964, 2021.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *Proc. International Conference on Machine Learning (ICML)*, volume 235, pp. 6116–6135, 2024.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Olivier Pietquin, and Mohammad Gheshlaghi Azar. Self-improving robust preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 4047–4068, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Miroslav Dudík, Katja Hofmann, Robert E Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In *Proc. Annual Conference on Learning Theory (CoLT)*, pp. 563–587. PMLR, 2015.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

Dhawal Gupta, Adam Fisch, Christoph Dann, and Alekh Agarwal. Mitigating preference hacking in policy optimization with pessimism. *arXiv preprint arXiv:2503.06810*, 2025.

Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. *Computational Optimization and Applications*, 61 (2):275–319, 2015.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason Lee, Wen Sun, Akshay Krishnamurthy, and Dylan Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 92647–92701, 2025.

Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pp. 824–832. PMLR, 2015.

Zimeng Li, Mudit Gaur, and Vaneet Aggarwal. Oracle-robust online alignment for large language models, 2026. URL https://arxiv.org/abs/2602.20457.

Yixin Liu, Argyris Oikonomou, Weiqiang Zheng, Yang Cai, and Arman Cohan. Comal: A convergent meta-algorithm for aligning llms with general preferences. In *Proc. International Conference on Learning Representations (ICLR)*, volume 2026, pp. 51383–51422, 2026.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 138663–138697, 2024.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 53728–53741, 2023.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free rlhf. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 37100–37137, 2024.

Ali Shirali, Arash Nasr-Esfahany, Abdullah Alomar, Parsa Mirtaheri, Rediet Abebe, and Ariel Procaccia. Direct alignment with heterogeneous preferences. *Advances in Neural Information Processing Systems*, 38:69858–69898, 2026.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: a roadmap to pluralistic alignment. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 3008–3021, 2020.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. Rspo: Regularized self-play alignment of large language models. *arXiv preprint arXiv:2503.00030*, 2025.

Daniil Tiapkin, Daniele Calandriello, Denis Belomestny, Eric Moulines, Alexey Naumov, Kashif Rasul, Michal Valko, and Pierre Menard. Accelerating nash learning from human feedback via mirror prox. *arXiv preprint arXiv:2505.19731*, 2025.

Mingzhi Wang, Chengdong Ma, Qizhi Chen, Linjian Meng, Yang Han, Jiancong Xiao, Zhaowei Zhang, Jing Huo, Weijie J Su, and Yaodong Yang. Magnetic preference optimization: Achieving last-iterate convergence for language model alignment. In *Proc. International Conference on Learning Representations (ICLR)*, 2025.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 16123–16150, 2025.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust LLM alignment via distributionally robust direct preference optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 81773–81807, 2024.

Yuheng Zhang, Dian Yu, Baolin Peng, Linfeng Song, Ye Tian, Mingyue Huo, Nan Jiang, Haitao Mi, and Dong Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

Runlong Zhou, Maryam Fazel, and Simon Shaolei Du. Extragradient preference optimization (EGPO): Beyond last-iterate convergence for nash learning from human feedback. In *Second Conference on Language Modeling*, 2025.

Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. PersonalLLM: Tailoring LLMs to individual preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2R7498e2Tx.

A Related Work

A.1 Reward modeling and direct preference optimization

Human comparisons support reward learning for reinforcement learning and language-model alignment (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). A standard pipeline fits scalar rewards and optimizes a policy with a reference penalty. The Bradley–Terry model (Bradley & Terry, 1952) links scalar rewards to comparison probabilities, imposing a scalar ordering that cannot represent arbitrary preference cycles.

Direct Preference Optimization (DPO) (Rafailov et al., 2023) uses the relation between rewards and KL-regularized optimal policies to replace reinforcement learning with a supervised preference loss, retaining the Bradley–Terry model. The more general ΨPO framework and IPO (Azar et al., 2024) examine how transformations of pairwise preferences change the optimization objective, distinguishing preference representation from training-loss design.

A.2 General preferences, von Neumann winners, and Nash alignment

In contextual dueling bandits, Dudík et al. (2015) introduce the von Neumann winner, a randomized policy that wins at least as often as it loses against every competitor, even when cycles preclude a universally preferred action. Nash Learning from Human Feedback (NLHF) (Munos et al., 2024) formulates a regularized preference game and introduces Nash Mirror Descent with last-iterate convergence guarantees. Swamy et al. (2024) likewise develop Self-Play Preference Optimization (SPO) from a minimax criterion for general preferences.

SPPO (Wu et al., 2024), online IPO and IPO-MD (Calandriello et al., 2024), and INPO (Zhang et al., 2024) develop iterative or self-play procedures for learning Nash policies from preference feedback. Complementing these algorithmic developments, Ye et al. (2024) establish statistical learning guarantees under general preference models. These works study equilibrium-based alignment under a fixed population preference relation.

A.3 Pessimistic Nash objectives and robust preference games

The closest objective-level precedent is Gupta et al. (2025), whose pessimistic Nash formulation minimizes over both an opponent policy and plausible preference models, with KL regularization on the policy pair. They show that unrestricted pessimism can favor poorly covered responses, motivating coverage-aware opponent restrictions and the practical P3O procedure. Self-Improving Robust Preference Optimization (SRPO) (Choi et al., 2025) addresses a different robustness source: it jointly optimizes generative and response-improvement policies to address task variation, with an offline supervised loss and an inference-time revision mechanism.

Our contribution is an exact convex–concave representation of the robust game over the full response simplex. A joint law over preference scenarios and opponent responses permits scenario-dependent opponents, while a conditional-KL penalty retains their reference regularization. The lift preserves the original inner worst-case value for every fixed leader policy, so its saddle-gap certificates apply to that original objective for both finite and continuous uncertainty families.

A.4 Heterogeneous preferences and pluralistic alignment

Heterogeneous-preference research distinguishes aggregate preference fitting from the representation of genuine user disagreement. Sorensen et al. (2024) advocate explicit treatment of plural human values; Shirali et al. (2026) characterize the annotator information needed for consistent direct alignment under heterogeneous user types; and Xiao et al. (2025) address preference collapse in KL-regularized reward optimization with a regularizer designed to match the preference distribution induced by their model.

MaxMin-RLHF (Chakraborty et al., 2024) optimizes worst-group reward, whereas Group Robust Preference Optimization (Ramesh et al., 2024) protects the worst-group preference-training loss. Group-specific comparison kernels can also serve as our scenarios, but the protected quantity is the regularized pairwise value against an adversarial opponent. These criteria need not agree even when

defined over the same groups. Scenarios may therefore encode either epistemic uncertainty about a population kernel or deliberate protection across groups; our guarantees apply to either supplied family, but their interpretation depends on this modeling choice.

A.5 Distributionally robust preference training

Distributionally robust preference training places uncertainty on observed preference-data distributions, following the broader robust-optimization paradigm (Bertsimas et al., 2011). Dr. DPO (Wu et al., 2025) distinguishes pointwise data quality from pairwise association errors when addressing noisy feedback, while Wasserstein DPO and KL DPO (Xu et al., 2025) address distribution shift through data-distribution ambiguity sets, with sample-complexity guarantees and scalable approximations. Closer to preference-model uncertainty, Li et al. (2026) study online alignment under pointwise misspecification of the preference oracle and, for log-linear policies, derive an exact decomposition of the robust objective into the nominal SAIL loss and an explicit sensitivity penalty.

Our robust Nash objective instead considers uncertain comparison kernels within a general-preference game in which an opponent also chooses its response distribution. Although the lift also uses probability measures, its scenario marginal ranges over all laws supported on the prescribed kernel family to represent worst-case scenario selection; it is not constrained to a divergence neighborhood of an empirical data distribution.

A.6 Optimization of regularized preference games

COMAL (Liu et al., 2026) obtains last-iterate convergence to a Nash policy of the original general-preference game through a convergent meta-algorithm. Magnetic Preference Optimization (Wang et al., 2025) combines magnetic mirror descent with an evolving reference, while Regularized Self-Play Policy Optimization (Tang et al., 2025) studies a family of regularizers and establishes convergence to the corresponding regularized equilibria. These approaches distinguish a fixed reference penalty, an algorithmic proximal anchor, and the equilibrium objective being optimized. For KL-regularized preference games, Zhou et al. (2025) establish linear last-iterate convergence with EGPO, while Tiapkin et al. (2025) analyze Nash Mirror-Prox and related proximal-point methods for parameterized self-play under regularity and approximation assumptions.

Our conditional penalty $\Psi(Q) = \mathrm{KL}(Q\|\alpha_Q \otimes r)$ regularizes the opponent responses but satisfies $\Psi(\alpha \otimes r) = 0$ for every scenario law α, leaving the marginal directions without strict convexity. Guarantees requiring curvature in both strategic distributions therefore do not directly apply. Adding a marginal penalty to the objective would generally change the value preserved by Theorem 3.1. We instead use Mirror-Prox, whose classical $O(1/T)$ ergodic analysis applies to Lipschitz monotone variational inequalities and smooth convex–concave saddle problems (Nemirovski, 2004). Composite Mirror-Prox (He et al., 2015) separates a smooth monotone component from a proximally tractable convex term: here the interaction is bilinear and both KL terms are handled proximally, avoiding a globally Lipschitz entropy-gradient requirement at the simplex boundary. Related multiplicative extragradient methods yield fast convergence in fully regularized matrix and Markov games (Cen et al., 2021).

For continuous uncertainty, a worst-case law may be singular relative to the diffuse initialization and have infinite KL. Following localization in continuum Hedge (Krichene et al., 2015), we approximate a point comparator by a distribution on a shrinking neighborhood, balancing payoff error against its logarithmic inverse-volume entropy cost. Relative-volume control and preservation of the comparator's conditional response policy yield a finite-entropy approximation without an additional conditional-KL error, extending the gap argument to the full continuous adversarial domain.

B Proofs of the main results

B.1 Notations

Joint laws and conditional regularization. Write $\mathcal{Z} := \Delta_m \times \mathcal{Q}$, where $\mathcal{Q} := \mathcal{P}(\mathcal{U} \times [m])$ is the set of Borel probability measures on $\mathcal{U} \times [m]$. For each $Q \in \mathcal{Q}$, let α_Q denote its scenario marginal

and choose a Borel conditional response kernel $\nu_Q(p) \in \Delta_m$ such that

$$Q(dp, j) = \alpha_Q(dp)\nu_{Q,j}(p).$$

Conversely, any probability law α on $\mathcal{U}$ and Borel kernel $\nu(p) \in \Delta_m$ define a joint law $Q \in \mathcal{Q}$ through

$$Q(dp, j) := \alpha(dp)\nu_j(p).$$

In particular, $\delta_p \otimes \nu \in \mathcal{Q}$ for every $p \in \mathcal{U}$ and $\nu \in \Delta_m$.

Define the conditional response regularizer

$$\Psi(Q) := \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Q(p)\|r\big)\, d\alpha_Q(p).$$

The conditional kernel is unique α_Q-almost everywhere, so $\Psi(Q)$ is independent of its values on α_Q-null sets.

In the finite case $\mathcal{U} = [K]$, identify Q with $q = (q_{kj}) \in \Delta_{Km}$ and set

$$\alpha_k := \sum_j q_{kj}, \qquad q_{kj} = \alpha_k \nu_{kj}.$$

For $\alpha_k > 0$, $\nu_{kj} := q_{kj}/\alpha_k$; for $\alpha_k = 0$, ν_k is immaterial. Then

$$\Psi(q) = \sum_k \alpha_k D_{\mathrm{KL}}(\nu_k\|r).$$

Analytical notation. For $z = (\pi, Q)$ and $u = (x, Y)$ in $\mathcal{Z}$, define

$$D_z(u) := D_{\mathrm{KL}}(x\|\pi) + D_{\mathrm{KL}}(Y\|Q).$$

Also define the composite regularizer

$$H(u) := \tau D_{\mathrm{KL}}(x\|r) + \tau\Psi(Y).$$

Relative entropy is $+\infty$ without absolute continuity, with $0\log(0/q) = 0$. For a bounded Borel direction (v, c), use

$$\langle (v, c), (x, Y)\rangle := v^\top x + \sum_{j=1}^m \int_{\mathcal{U}} c_j(p)\, Y(dp, j),$$

extended linearly to increments and signed joint measures. We use the full total variation norm

$$\|M\|_{\mathrm{var}} := |M|(\mathcal{U} \times [m]).$$

Standing bounds. Let

$$r_{\min} := \min_j r_j, \qquad R := \log(1/r_{\min}).$$

For every $v \in \Delta_m$,

$$0 \le D_{\mathrm{KL}}(v\|r) = \sum_j v_j \log v_j - \sum_j v_j \log r_j \le R,$$

because the first sum is nonpositive and $-\log r_j \le R$. Consequently,

$$0 \le \Psi(Y) \le R, \qquad 0 \le H(u) \le 2\tau R.$$

B.2 Proof of Theorem 3.1: exact joint reformulation

Proof. **Fixed-policy exactness.** Fix $\pi \in \Delta_m$. As $J(\pi, \nu; A(p))$ is continuous in (ν, p) on the compact set $\mathcal{U} \times \Delta_m$, it has a minimizing pair $(p^\star, \nu^\star)$. By the definitions of $s(Q)$ and $\Psi(Q)$,

$$\pi^\top s(Q) = \int_{\mathcal{U}} \pi^\top A(p)\nu_Q(p)\, d\alpha_Q(p), \qquad \Psi(Q) = \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Q(p)\|r\big)\, d\alpha_Q(p).$$

Since $\alpha_Q(\mathcal{U}) = 1$, the leader regularizer can also be placed inside the integral. For every $Q \in \mathcal{Q}$,

$$\begin{aligned} L(\pi, Q) &= \int_{\mathcal{U}} J\big(\pi, \nu_Q(p); A(p)\big)\, d\alpha_Q(p) \\ &\geq \min_{p \in \mathcal{U},\, \nu \in \Delta_m} J(\pi, \nu; A(p)) = g(\pi). \end{aligned} \tag{18}$$

Conversely, $Q^\star_\pi := \delta_{p^\star} \otimes \nu^\star$ attains equality. Hence

$$\min_{Q \in \mathcal{Q}} L(\pi, Q) \geq g(\pi) = L(\pi, Q^\star_\pi) \geq \min_{Q \in \mathcal{Q}} L(\pi, Q), \tag{19}$$

i.e., $\min_{Q \in \mathcal{Q}} L(\pi, Q) = g(\pi)$.

Convex–concave structure. Because $\mathcal{U} \times [m]$ is compact metric, $\mathcal{Q} = \mathcal{P}(\mathcal{U} \times [m])$ is nonempty, convex, and compact in the weak topology, even if $\mathcal{U}$ is not convex. Each coordinate of $s(Q)$ integrates the bounded continuous function $(p, j) \mapsto A_{ij}(p)$, so s is affine and weakly continuous. The product law $\alpha_Q \otimes r$ has the same scenario marginal α_Q as Q and has conditional response law r at every p. Applying Lemma C.1 therefore gives

$$\begin{aligned} D_{\mathrm{KL}}(Q \| \alpha_Q \otimes r) &= D_{\mathrm{KL}}(\alpha_Q \| \alpha_Q) + \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Q(p) \| r\big)\, d\alpha_Q(p) \\ &= \Psi(Q), \end{aligned} \tag{20}$$

where the last equality is the definition of Ψ from Appendix B.1. The marginal map $Q \mapsto \alpha_Q$ is affine and weakly continuous, as is $\alpha \mapsto \alpha \otimes r$, since for every bounded continuous f the function $p \mapsto \sum_j r_j f(p, j)$ is continuous. Indeed, for $Q_\theta = \theta Q_1 + (1 - \theta) Q_2$,

$$\alpha_{Q_\theta} \otimes r = \theta(\alpha_{Q_1} \otimes r) + (1 - \theta)(\alpha_{Q_2} \otimes r).$$

Joint convexity of relative entropy therefore yields

$$\begin{aligned} \Psi(Q_\theta) &= D_{\mathrm{KL}}(Q_\theta \| \alpha_{Q_\theta} \otimes r) \\ &\leq \theta D_{\mathrm{KL}}(Q_1 \| \alpha_{Q_1} \otimes r) + (1 - \theta) D_{\mathrm{KL}}(Q_2 \| \alpha_{Q_2} \otimes r) \\ &= \theta \Psi(Q_1) + (1 - \theta) \Psi(Q_2). \end{aligned}$$

Likewise, if $Q_n \Rightarrow Q$, weak continuity of the marginal map implies

$$\alpha_{Q_n} \Rightarrow \alpha_Q, \qquad \alpha_{Q_n} \otimes r \Rightarrow \alpha_Q \otimes r.$$

Joint weak lower semicontinuity of relative entropy then gives

$$\begin{aligned} \Psi(Q) &= D_{\mathrm{KL}}(Q \| \alpha_Q \otimes r) \\ &\leq \liminf_{n \to \infty} D_{\mathrm{KL}}(Q_n \| \alpha_{Q_n} \otimes r) = \liminf_{n \to \infty} \Psi(Q_n). \end{aligned}$$

Hence Ψ is convex and weakly lower semicontinuous.

Consequently, for every fixed Q,

$$L(\pi, Q) = \pi^\top s(Q) - \tau D_{\mathrm{KL}}(\pi \| r) + \tau \Psi(Q)$$

is continuous and concave in π: the first term is linear, while $-D_{\mathrm{KL}}(\pi \| r)$ is continuous and concave because r has full support. For every fixed π, $Q \mapsto \pi^\top s(Q)$ is affine and weakly continuous, while Ψ is convex and weakly lower semicontinuous. Hence $L(\pi, \cdot)$ is convex and weakly lower semicontinuous on $\mathcal{Q}$.

Minimax equality and saddle points. The hypotheses of Sion's minimax theorem are satisfied: Δ_m and $\mathcal{Q}$ are convex, with Δ_m compact and $\mathcal{Q}$ compact in the weak topology; moreover, $L(\cdot, Q)$ is continuous and concave for every Q, while $L(\pi, \cdot)$ is convex and weakly lower semicontinuous for every π. Hence Sion's minimax theorem (Sion, 1958) gives

$$\max_{\pi \in \Delta_m} \min_{Q \in \mathcal{Q}} L(\pi, Q) = \min_{Q \in \mathcal{Q}} \max_{\pi \in \Delta_m} L(\pi, Q). \tag{21}$$

Equation 19 identifies the left-hand side with $V^\star$.

The function g is continuous because it is the minimum of the continuous payoff J over the fixed compact set $\mathcal{U} \times \Delta_m$, and hence it attains its maximum. Likewise, $Q \mapsto \max_\pi L(\pi, Q)$ is weakly lower semicontinuous and attains its minimum on compact $\mathcal{Q}$. Let $\pi^\star$ and $Q^\star$ be any corresponding outer optimizers. Then

$$L(\pi, Q^\star) \leq V^\star = L(\pi^\star, Q^\star) \leq L(\pi^\star, Q), \qquad \pi \in \Delta_m, \quad Q \in \mathcal{Q}.$$

Thus $(\pi^\star, Q^\star)$ is a saddle point, completing the proof. □

B.3 Closed-form joint entropic proximal update

Proof. Set $a := 1 + \eta\tau$. We derive the leader update, the conditional response updates, and finally the scenario weights.

Leader update. Define

$$w_i := (\pi_i^\circ)^{1/a} r_i^{\eta\tau/a} e^{\eta s_i/a}, \qquad W := \sum_i w_i, \qquad \pi_i^+ := w_i/W.$$

The leader's proximal objective, written as a minimization, is

$$\begin{aligned}\mathcal{J}_\pi(\pi) &:= D_{\mathrm{KL}}(\pi\|\pi^\circ) + \eta\tau D_{\mathrm{KL}}(\pi\|r) - \eta s^\top \pi \\ &= \sum_i \pi_i \log\frac{\pi_i}{\pi_i^\circ} + \eta\tau \sum_i \pi_i \log\frac{\pi_i}{r_i} - \eta \sum_i \pi_i s_i \\ &= \sum_i \pi_i \left[a \log \pi_i - \log \pi_i^\circ - \eta\tau \log r_i - \eta s_i\right]. \end{aligned} \tag{22}$$

By the definition of w_i,

$$a \log w_i = \log \pi_i^\circ + \eta\tau \log r_i + \eta s_i,$$

so

$$\begin{aligned}\mathcal{J}_\pi(\pi) &= a \sum_i \pi_i \log\frac{\pi_i}{w_i} \\ &= a \sum_i \pi_i \log\frac{\pi_i}{W\pi_i^+} \\ &= a D_{\mathrm{KL}}(\pi\|\pi^+) - a \log W, \end{aligned} \tag{23}$$

where the last equality uses $w_i = W\pi_i^+$ and $\sum_i \pi_i = 1$. Since $a > 0$ and $D_{\mathrm{KL}}(\pi\|\pi^+) \geq 0$, with equality if and only if $\pi = \pi^+$, the vector π^+ is the unique minimizer.

Conditional response updates. Write

$$q_{kj} = \alpha_k \nu_{kj}, \qquad q_{kj}^\circ = \alpha_k^\circ \nu_{kj}^\circ.$$

For rows with $\alpha_k = 0$, the conditional law ν_k is immaterial and all terms weighted by α_k are understood as zero. Expanding the joint KL gives

$$\begin{aligned} D_{\mathrm{KL}}(q\|q^\circ) &= \sum_{k,j} \alpha_k \nu_{kj} \log\frac{\alpha_k \nu_{kj}}{\alpha_k^\circ \nu_{kj}^\circ} \\ &= \sum_k \alpha_k \log\frac{\alpha_k}{\alpha_k^\circ} + \sum_k \alpha_k \sum_j \nu_{kj} \log\frac{\nu_{kj}}{\nu_{kj}^\circ} \\ &= D_{\mathrm{KL}}(\alpha\|\alpha^\circ) + \sum_k \alpha_k D_{\mathrm{KL}}(\nu_k\|\nu_k^\circ). \end{aligned}$$

Moreover, by the finite-scenario form of Ψ and the factorization $q_{kj} = \alpha_k \nu_{kj}$,

$$\Psi(q) = \sum_k \alpha_k D_{\mathrm{KL}}(\nu_k\|r), \qquad \langle c, q\rangle = \sum_k \alpha_k c_k^\top \nu_k.$$

Hence

$$\begin{aligned}\mathcal{J}_q(q) &:= D_{\mathrm{KL}}(q\|q^\circ) + \eta\tau\Psi(q) + \eta\langle c, q\rangle \\ &= D_{\mathrm{KL}}(\alpha\|\alpha^\circ) + \sum_k \alpha_k \Phi_k(\nu_k), \end{aligned} \tag{24}$$

where

$$\Phi_k(\nu) := D_{\mathrm{KL}}(\nu\|\nu_k^\circ) + \eta\tau D_{\mathrm{KL}}(\nu\|r) + \eta c_k^\top \nu.$$

Now set

$$b_{kj} := (\nu^\circ_{kj})^{1/a} r_j^{\eta\tau/a} e^{-\eta c_{kj}/a}, \qquad Z_k := \sum_j b_{kj}, \qquad \nu^+_{kj} := \frac{b_{kj}}{Z_k}.$$

Expanding Φ_k gives

$$\begin{aligned}\Phi_k(\nu) &= \sum_j \nu_j \left[\log \frac{\nu_j}{\nu^\circ_{kj}} + \eta\tau \log \frac{\nu_j}{r_j} + \eta c_{kj}\right] \\ &= \sum_j \nu_j \left[a \log \nu_j - \log \nu^\circ_{kj} - \eta\tau \log r_j + \eta c_{kj}\right].\end{aligned}$$

Since

$$a \log b_{kj} = \log \nu^\circ_{kj} + \eta\tau \log r_j - \eta c_{kj},$$

we obtain

$$\begin{aligned}\Phi_k(\nu) &= a \sum_j \nu_j \log \frac{\nu_j}{b_{kj}} \\ &= a \sum_j \nu_j \log \frac{\nu_j}{Z_k \nu^+_{kj}} \\ &= a D_{\mathrm{KL}}(\nu \| \nu^+_k) - a \log Z_k. \end{aligned} \tag{25}$$

Since $D_{\mathrm{KL}}(\nu\|\nu^+_k) \geq 0$, with equality only at $\nu = \nu^+_k$,

$$\min_{\nu \in \Delta_m} \Phi_k(\nu) = -a \log Z_k,$$

and ν^+_k is the unique conditional minimizer.

Scenario weights. Substituting the optimized conditional values $\min_\nu \Phi_k(\nu) = -a \log Z_k$ into equation 24 leaves

$$\min_{\alpha \in \Delta_K} \left\{ D_{\mathrm{KL}}(\alpha \| \alpha^\circ) - a \sum_k \alpha_k \log Z_k \right\}. \tag{26}$$

Define

$$C := \sum_\ell \alpha^\circ_\ell Z^a_\ell, \qquad \alpha^+_k := \frac{\alpha^\circ_k Z^a_k}{C}.$$

Then

$$\log \alpha^+_k = \log \alpha^\circ_k + a \log Z_k - \log C.$$

Therefore,

$$\begin{aligned} & D_{\mathrm{KL}}(\alpha \| \alpha^\circ) - a \sum_k \alpha_k \log Z_k \\ &\quad = \sum_k \alpha_k \left(\log \frac{\alpha_k}{\alpha^\circ_k} - a \log Z_k \right) \\ &\quad = \sum_k \alpha_k \log \frac{\alpha_k}{C \alpha^+_k} \\ &\quad = D_{\mathrm{KL}}(\alpha \| \alpha^+) - \log C, \end{aligned} \tag{27}$$

where the last step uses $\sum_k \alpha_k = 1$. Hence α^+ is the unique minimizing scenario marginal. Combining the two joint completions yields

$$\mathcal{J}_q(q) = -\log C + D_{\mathrm{KL}}(\alpha \| \alpha^+) + a \sum_k \alpha_k D_{\mathrm{KL}}(\nu_k \| \nu^+_k).$$

The nonnegative KL terms vanish exactly when $q = q^+$, where $q^+_{kj} := \alpha^+_k \nu^+_{kj}$. Together with equation 23, this proves the formulas and uniqueness in equation 11. Positive centers and finite scores and costs make all weights positive and all normalizers finite. The factor $Z_k^{1+\eta\tau}$ follows from the optimized conditional value $-(1+\eta\tau)\log Z_k$. □

Algorithm 1 Exact Joint Entropic Mirror-Prox: Finite Scenarios

Require: Matrices $\{A^k\}_{k=1}^K$, full-support reference r, $\tau > 0$, fixed $0 < \eta \le 1$, integer $T \ge 1$

$z_1 = (\pi_1, q_1) \leftarrow \big(r, (r_j/K)_{k,j}\big)$; $\bar{z}_0 \leftarrow 0$
for $t = 1, \dots, T$ **do**
  $s_t \leftarrow s(q_t)$; $c_t \leftarrow c(\pi_t)$
  $\widetilde{z}_t = (\widetilde{\pi}_t, \widetilde{q}_t) \leftarrow \mathrm{JProx}_\eta(z_t; s_t, c_t)$ ▷ predictor
  $\widetilde{s}_t \leftarrow s(\widetilde{q}_t)$; $\widetilde{c}_t \leftarrow c(\widetilde{\pi}_t)$
  $z_{t+1} \leftarrow \mathrm{JProx}_\eta(z_t; \widetilde{s}_t, \widetilde{c}_t)$ ▷ corrector; same center
  $\bar{z}_t \leftarrow \bar{z}_{t-1} + (\widetilde{z}_t - \bar{z}_{t-1})/t$
end for
return $(\bar{\pi}_T, \bar{q}_T) = \bar{z}_T$

B.4 Proof of Theorem 4.2: finite-scenario convergence

Proof. Algorithm 1 initializes

$$\pi_1 = r, \qquad q_{1,kj} = \frac{r_j}{K},$$

which are strictly positive because r has full support. By Proposition 4.1, a positive proximal center together with finite scores and costs produces strictly positive leader and joint outputs. Hence every predictor, corrector, and subsequent center remains strictly positive.

Since the domains are finite, every $u = (x, y) \in \Delta_m \times \Delta_{Km}$ therefore satisfies

$$D_{z_t}(u) = D_{\mathrm{KL}}(x\|\pi_t) + D_{\mathrm{KL}}(y\|q_t) < \infty$$

at every finite iteration. For unit-range kernels, Lemma C.3 permits $L_F = 1$. Thus $0 < \eta \le 1$ implies $\eta L_F \le 1$, so the one-step estimate is applicable at every round. Applying Lemma C.3 at round t gives, for every fixed $u = (x, y) \in \Delta_m \times \Delta_{Km}$,

$$\eta\big[L(x, \widetilde{q}_t) - L(\widetilde{\pi}_t, y)\big] \le D_{z_t}(u) - D_{z_{t+1}}(u).$$

Summing this inequality over $t = 1, \dots, T$ telescopes the divergence terms:

$$\eta \sum_{t=1}^{T} \big[L(x, \widetilde{q}_t) - L(\widetilde{\pi}_t, y)\big] \le D_{z_1}(u) - D_{z_{T+1}}(u).$$

By definition,

$$\bar{\pi}_T = \frac{1}{T}\sum_{t=1}^{T} \widetilde{\pi}_t, \qquad \bar{q}_T = \frac{1}{T}\sum_{t=1}^{T} \widetilde{q}_t.$$

For fixed x, the map $q \mapsto L(x, q)$ is convex, so Jensen's inequality gives

$$L(x, \bar{q}_T) \le \frac{1}{T}\sum_{t=1}^{T} L(x, \widetilde{q}_t).$$

For fixed y, the map $\pi \mapsto L(\pi, y)$ is concave, hence

$$L(\bar{\pi}_T, y) \ge \frac{1}{T}\sum_{t=1}^{T} L(\widetilde{\pi}_t, y).$$

Subtracting these two inequalities and using the telescoping bound yields

$$\begin{aligned} L(x, \bar{q}_T) - L(\bar{\pi}_T, y) &\le \frac{1}{T}\sum_{t=1}^{T} \big[L(x, \widetilde{q}_t) - L(\widetilde{\pi}_t, y)\big] \\ &\le \frac{D_{z_1}(u) - D_{z_{T+1}}(u)}{\eta T} \\ &\le \frac{D_{z_1}(u)}{\eta T}, \end{aligned} \tag{28}$$

where the last inequality uses $D_{z_{T+1}}(u) \geq 0$. It remains to bound the initial divergence. Since $z_1 = (r, q_1)$,

$$D_{z_1}(u) = D_{\mathrm{KL}}(x\|r) + D_{\mathrm{KL}}(y\|q_1).$$

For the leader component,

$$\begin{aligned} D_{\mathrm{KL}}(x\|r) &= \sum_i x_i \log x_i - \sum_i x_i \log r_i \\ &\leq R, \end{aligned}$$

because $\sum_i x_i \log x_i \leq 0$ and $-\log r_i \leq R$.

For the joint component, using $q_{1,kj} = r_j/K$,

$$\begin{aligned} D_{\mathrm{KL}}(y\|q_1) &= \sum_{k,j} y_{kj} \log \frac{y_{kj}}{r_j/K} \\ &= \sum_{k,j} y_{kj} \log y_{kj} + \log K - \sum_{k,j} y_{kj} \log r_j \\ &\leq \log K + R. \end{aligned}$$

These bounds are sharp. Let $i^\star \in \arg\min_i r_i$ and $j^\star \in \arg\min_j r_j$. Taking $x = e_{i^\star}$ gives $D_{\mathrm{KL}}(x\|r) = R$, while for any scenario $k^\star$, taking $y = e_{(k^\star, j^\star)}$ gives

$$D_{\mathrm{KL}}(y\|q_1) = \log K + R.$$

Since x and y are independent comparison variables,

$$\sup_{x \in \Delta_m,\, y \in \Delta_{Km}} D_{z_1}(x, y) = R + (\log K + R) = \log K + 2R. \tag{29}$$

Taking the supremum over the independent variables x and y gives

$$\begin{aligned} \sup_{x,y} \big\{ L(x, \bar{q}_T) - L(\bar{\pi}_T, y) \big\} &= \max_{x \in \Delta_m} L(x, \bar{q}_T) - \min_{y \in \Delta_{Km}} L(\bar{\pi}_T, y) \\ &= \mathrm{Gap}(\bar{\pi}_T, \bar{q}_T), \end{aligned}$$

where the second equality uses the exact finite joint reformulation $\min_y L(\bar{\pi}_T, y) = g(\bar{\pi}_T)$. Combining this identity with equation 28 and equation 29 yields

$$\mathrm{Gap}(\bar{\pi}_T, \bar{q}_T) \leq \frac{\log K + 2R}{\eta T}.$$

Moreover,

$$g(\bar{\pi}_T) = \min_y L(\bar{\pi}_T, y) \leq L(\bar{\pi}_T, \bar{q}_T) \leq \max_x L(x, \bar{q}_T),$$

so $\mathrm{Gap}(\bar{\pi}_T, \bar{q}_T) \geq 0$. □

B.5 Proof of Theorem 5.1: the continuous full-domain gap

Proof. Let

$$E := \mathrm{aff}(\mathcal{U}), \qquad d_{\mathcal{U}} := \dim E,$$

and let $\lambda_{\mathcal{U}}$ denote the normalized $d_{\mathcal{U}}$-dimensional Lebesgue measure on $\mathcal{U}$ relative to E, interpreted as the Dirac law when $d_{\mathcal{U}} = 0$. Fix $T \geq 1$ and write

$$Q_1 := \lambda_{\mathcal{U}} \otimes r.$$

Admissible comparators. Under exact moment access, Algorithm 2 uses the prox of Lemma C.2. Consider either proximal output $w = (\pi_w, Q_w)$ from a center z and a comparator $u = (x, Y)$ with $D_z(u) < \infty$. Let $\mathcal{J}$ denote the corresponding proximal objective in Lemma C.2. Its completion identity gives

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(w) = a D_{\mathrm{KL}}(x\|\pi_w) + D_{\mathrm{KL}}(\alpha_Y\|\alpha_{Q_w}) \\ + a \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Y(p)\|\nu_{Q_w}(p)\big)\, d\alpha_Y(p). \end{aligned}$$

Algorithm 2 Exact Joint Entropic Mirror-Prox: Compact-Convex Uncertainty

Require: $\mathcal{U}, A, r, \tau$ as above, fixed $0 < \eta \le 1$, integer $T \ge 1$, and exact moment access

$z_1 = (\pi_1, Q_1) \leftarrow (r, \lambda_{\mathcal{U}} \otimes r)$; $\bar{z}_0 \leftarrow 0$
$s_1 \leftarrow \int_{\mathcal{U}} A(p) r \, d\lambda_{\mathcal{U}}(p)$
for $t = 1, \dots, T$ **do**
  $(\widetilde{z}_t, \widetilde{s}_t) \leftarrow \textsc{MeasureProx}(z_t; s_t, A(\cdot)^\top \pi_t)$ ▷ predictor
  $(z_{t+1}, s_{t+1}) \leftarrow \textsc{MeasureProx}(z_t; \widetilde{s}_t, A(\cdot)^\top \widetilde{\pi}_t)$ ▷ same center
  $\bar{z}_t \leftarrow \bar{z}_{t-1} + (\widetilde{z}_t - \bar{z}_{t-1})/t$
end for
return $(\bar{\pi}_T, \bar{Q}_T) = \bar{z}_T$

function $\textsc{MeasureProx}(z^\circ; s, c)$
  Compute π^+ and pointwise $b(p), Z(p), \nu^+(p)$ using equation 11
  $C \leftarrow \int_{\mathcal{U}} Z(p)^a \, d\alpha^\circ(p)$; $\alpha^+(dp) \leftarrow C^{-1} Z(p)^a \alpha^\circ(dp)$
  $Q^+(dp, j) \leftarrow \alpha^+(dp) \nu_j^+(p)$; $s^+ \leftarrow \int_{\mathcal{U}} A(p) \nu^+(p) \, d\alpha^+(p)$
  return $\big((\pi^+, Q^+), s^+\big)$
end function

By the KL chain rule,

$$D_w(u) = D_{\mathrm{KL}}(x \| \pi_w) + D_{\mathrm{KL}}(\alpha_Y \| \alpha_{Q_w}) + \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Y(p) \| \nu_{Q_w}(p)\big) \, d\alpha_Y(p).$$

Since $a = 1 + \eta\tau \ge 1$,

$$\mathcal{J}(u) - \mathcal{J}(w) \ge D_w(u).$$

The left-hand side is finite because $D_z(u) < \infty$, the proximal direction is bounded, and H is bounded. Hence $D_w(u) < \infty$. Applying this argument successively to the predictor and corrector shows that every comparator with $D_{z_1}(u) < \infty$ remains admissible at every finite iteration.

Comparison bound. For $0 < \eta \le 1$, Lemma C.3 applies at every round. For every fixed $u = (x, Y)$ with $D_{z_1}(u) < \infty$, it gives

$$\eta\big[L(x, \widetilde{Q}_t) - L(\widetilde{\pi}_t, Y)\big] \le D_{z_t}(u) - D_{z_{t+1}}(u).$$

Summing over $t = 1, \dots, T$ telescopes the divergence terms:

$$\eta \sum_{t=1}^{T} \big[L(x, \widetilde{Q}_t) - L(\widetilde{\pi}_t, Y)\big] \le D_{z_1}(u) - D_{z_{T+1}}(u).$$

Since

$$\bar{\pi}_T = \frac{1}{T} \sum_{t=1}^{T} \widetilde{\pi}_t, \qquad \bar{Q}_T = \frac{1}{T} \sum_{t=1}^{T} \widetilde{Q}_t,$$

convexity of $L(x, \cdot)$ and concavity of $L(\cdot, Y)$ give

$$L(x, \bar{Q}_T) \le \frac{1}{T} \sum_{t=1}^{T} L(x, \widetilde{Q}_t), \qquad L(\bar{\pi}_T, Y) \ge \frac{1}{T} \sum_{t=1}^{T} L(\widetilde{\pi}_t, Y).$$

Combining these inequalities yields

$$\begin{aligned} L(x, \bar{Q}_T) - L(\bar{\pi}_T, Y) &\le \frac{1}{T} \sum_{t=1}^{T} \big[L(x, \widetilde{Q}_t) - L(\widetilde{\pi}_t, Y)\big] \\ &\le \frac{D_{z_1}(u) - D_{z_{T+1}}(u)}{\eta T} \\ &\le \frac{D_{\mathrm{KL}}(x \| r) + D_{\mathrm{KL}}(Y \| Q_1)}{\eta T}, \end{aligned} \tag{30}$$

where the last step uses $z_1 = (r, Q_1)$ and $D_{z_{T+1}}(u) \geq 0$. Since the trajectory is independent of the comparator u, the bound holds simultaneously for every u with $D_{z_1}(u) < \infty$. Hence an admissible comparator chosen after observing $(\bar{\pi}_T, \bar{Q}_T)$ may be substituted below.

Localization. Suppose $d_{\mathcal{U}} \geq 1$. By the definition of $d_{\mathcal{U}}$, the set $\mathcal{U}$ contains $d_{\mathcal{U}} + 1$ affinely independent points in E. Their convex hull is a full-dimensional simplex contained in $\mathcal{U}$, so $\mathcal{U}$ has strictly positive relative $d_{\mathcal{U}}$-dimensional volume. Compactness makes this volume finite.

Moreover, $\lambda_{\mathcal{U}}$ has support $\mathcal{U}$. Indeed, a relative-interior ball contained in $\mathcal{U}$ can be contracted toward any $p \in \mathcal{U}$; for a sufficiently small contraction, its image lies in any prescribed relative neighborhood of p and has positive relative volume.

Choose a worst pair $(p^\star, \nu^\star)$ for $\bar{\pi}_T$ by Theorem 3.1. For $0 < \sigma \leq 1$, define the contraction map

$$\Gamma_\sigma(p) := (1 - \sigma)p^\star + \sigma p.$$

For a measurable map Γ and a probability law α, write $\Gamma_{\#}\alpha$ for the pushforward law,

$$(\Gamma_{\#}\alpha)(B) := \alpha(\Gamma^{-1}(B)).$$

Set

$$\alpha_\sigma := (\Gamma_\sigma)_{\#}\lambda_{\mathcal{U}}, \qquad Y_\sigma := \alpha_\sigma \otimes \nu^\star.$$

Convexity gives $\Gamma_\sigma(\mathcal{U}) \subseteq \mathcal{U}$. The map Γ_σ is an injective affine homothety on E whose linear part is multiplication by σ. Hence it scales $d_{\mathcal{U}}$-dimensional relative volume by $\sigma^{d_{\mathcal{U}}}$. For every Borel set $B \subseteq \mathcal{U}$,

$$\begin{aligned}\alpha_\sigma(B) &= \lambda_{\mathcal{U}}\big(\Gamma_\sigma^{-1}(B)\big) \\ &= \sigma^{-d_{\mathcal{U}}}\lambda_{\mathcal{U}}\big(B \cap \Gamma_\sigma(\mathcal{U})\big).\end{aligned}$$

Therefore

$$\frac{d\alpha_\sigma}{d\lambda_{\mathcal{U}}} = \sigma^{-d_{\mathcal{U}}}\mathbf{1}_{\Gamma_\sigma(\mathcal{U})}, \qquad \lambda_{\mathcal{U}}\big(\Gamma_\sigma(\mathcal{U})\big) = \sigma^{d_{\mathcal{U}}}. \tag{31}$$

Using equation 31,

$$\begin{aligned}D_{\mathrm{KL}}(\alpha_\sigma \| \lambda_{\mathcal{U}}) &= \int_{\mathcal{U}} \log \frac{d\alpha_\sigma}{d\lambda_{\mathcal{U}}}\, d\alpha_\sigma \\ &= d_{\mathcal{U}} \log(1/\sigma),\end{aligned}$$

because α_σ is supported on $\Gamma_\sigma(\mathcal{U})$. Since

$$Y_\sigma = \alpha_\sigma \otimes \nu^\star, \qquad Q_1 = \lambda_{\mathcal{U}} \otimes r,$$

Lemma C.1 then gives

$$\begin{aligned}D_{\mathrm{KL}}(Y_\sigma \| Q_1) &= D_{\mathrm{KL}}(\alpha_\sigma \| \lambda_{\mathcal{U}}) + D_{\mathrm{KL}}(\nu^\star \| r) \\ &= d_{\mathcal{U}} \log(1/\sigma) + D_{\mathrm{KL}}(\nu^\star \| r).\end{aligned} \tag{32}$$

In particular, Y_σ has finite initial KL. Thus

$$\begin{aligned}D_{z_1}(x, Y_\sigma) &= D_{\mathrm{KL}}(x \| r) + D_{\mathrm{KL}}(Y_\sigma \| Q_1) \\ &\leq 2R + d_{\mathcal{U}} \log(1/\sigma).\end{aligned}$$

Substituting this finite comparator into equation 30 gives

$$L(x, \bar{Q}_T) - L(\bar{\pi}_T, Y_\sigma) \leq \frac{2R + d_{\mathcal{U}} \log(1/\sigma)}{\eta T}. \tag{33}$$

We next compare the localized law with the exact worst pair. Since $(p^\star, \nu^\star)$ minimizes $J(\bar{\pi}_T, \nu; A(p))$ over (p, ν),

$$g(\bar{\pi}_T) = J(\bar{\pi}_T, \nu^\star; A(p^\star)).$$

The law Y_σ keeps the same leader $\bar{\pi}_T$ and the same conditional response $\nu^\star$, so the two KL regularizers are identical in $L(\bar{\pi}_T, Y_\sigma)$ and $g(\bar{\pi}_T)$ and cancel.

By the definition of pushforward,

$$\int_{\mathcal{U}} f(p)\, d\alpha_\sigma(p) = \int_{\mathcal{U}} f(\Gamma_\sigma(p))\, d\lambda_{\mathcal{U}}(p).$$

Since A is affine,

$$A(\Gamma_\sigma(p)) = (1-\sigma)A(p^\star) + \sigma A(p).$$

Consequently,

$$\begin{aligned} L(\bar{\pi}_T, Y_\sigma) - g(\bar{\pi}_T) &= \int_{\mathcal{U}} \bar{\pi}_T^\top [A(\Gamma_\sigma(p)) - A(p^\star)]\nu^\star \, d\lambda_{\mathcal{U}}(p) \\ &= \sigma \int_{\mathcal{U}} \bar{\pi}_T^\top [A(p) - A(p^\star)]\nu^\star \, d\lambda_{\mathcal{U}}(p). \end{aligned} \tag{34}$$

For every $p \in \mathcal{U}$, worst-pair optimality gives

$$J(\bar{\pi}_T, \nu^\star; A(p)) \geq J(\bar{\pi}_T, \nu^\star; A(p^\star)).$$

The KL terms are identical on the two sides, so

$$\bar{\pi}_T^\top [A(p) - A(p^\star)]\nu^\star \geq 0.$$

Moreover, because $A(p) \in [0,1]^{m\times m}$ and both $\bar{\pi}_T$ and $\nu^\star$ are probability vectors, each bilinear payoff lies in $[0,1]$. Hence

$$0 \leq \bar{\pi}_T^\top [A(p) - A(p^\star)]\nu^\star \leq 1.$$

Therefore

$$0 \leq L(\bar{\pi}_T, Y_\sigma) - g(\bar{\pi}_T) \leq \sigma. \tag{35}$$

By definition,

$$\mathrm{Gap}(\bar{\pi}_T, \bar{Q}_T) = \max_{x\in\Delta_m} L(x, \bar{Q}_T) - g(\bar{\pi}_T).$$

Adding and subtracting $L(\bar{\pi}_T, Y_\sigma)$ gives

$$\begin{aligned} \mathrm{Gap}(\bar{\pi}_T, \bar{Q}_T) &= \Big[\max_{x\in\Delta_m} L(x, \bar{Q}_T) - L(\bar{\pi}_T, Y_\sigma)\Big] \\ &\quad + \Big[L(\bar{\pi}_T, Y_\sigma) - g(\bar{\pi}_T)\Big] \\ &\leq \frac{2R + d_{\mathcal{U}} \log(1/\sigma)}{\eta T} + \sigma, \end{aligned}$$

where the last inequality follows from equation 33 and equation 35. The localized law Y_σ is used only as a finite-KL comparator; the robust value $g(\bar{\pi}_T)$ still minimizes over the full joint-law domain $\mathcal{Q}$, including singular laws.

Finally,

$$g(\bar{\pi}_T) = \min_Y L(\bar{\pi}_T, Y) \leq L(\bar{\pi}_T, \bar{Q}_T) \leq \max_x L(x, \bar{Q}_T),$$

so the gap is nonnegative. Choosing $\eta = 1/2$ and $\sigma = 1/T$ gives

$$\frac{4R + 2d_{\mathcal{U}} \log T + 1}{T}.$$

Alternatively, with $\eta = 1/2$ and $\sigma = \varepsilon/2$,

$$\mathrm{Gap}(\bar{\pi}_T, \bar{Q}_T) \leq \frac{4R + 2d_{\mathcal{U}} \log(2/\varepsilon)}{T} + \frac{\varepsilon}{2}.$$

Thus it suffices to make the first term at most $\varepsilon/2$, which holds whenever

$$T \geq \left\lceil \frac{8R + 4d_{\mathcal{U}} \log(2/\varepsilon)}{\varepsilon} \right\rceil. \tag{36}$$

Singleton case. If $d_{\mathcal{U}} = 0$, then the nonempty set $\mathcal{U}$ has zero-dimensional affine hull and hence $\mathcal{U} = \{p_0\}$. Every joint law on $\mathcal{U} \times [m]$ therefore has the form $Y = \delta_{p_0} \otimes \nu$. Thus $D_{z_1}(x, Y) = D_{\mathrm{KL}}(x\|r) + D_{\mathrm{KL}}(\nu\|r) \leq 2R$. Taking the supremum over x, ν in equation 30 directly gives

$$0 \leq \mathrm{Gap}(\bar{\pi}_T, \bar{Q}_T) \leq \frac{2R}{\eta T}.$$

□

C AUXILIARY LEMMAS

C.1 KL CHAIN RULE FOR JOINT LAWS

Lemma C.1 (Joint KL chain rule)**.** *Let*

$$Y(dp, j) = \alpha(dp)\nu_j(p), \qquad Q(dp, j) = \beta(dp)\mu_j(p)$$

be probability laws on $\mathcal{U} \times [m]$. Then, in the extended nonnegative reals,

$$D_{\mathrm{KL}}(Y\|Q) = D_{\mathrm{KL}}(\alpha\|\beta) + \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu(p)\|\mu(p)\big)\, d\alpha(p). \tag{37}$$

Proof. If $\alpha \not\ll \beta$, then $Y \not\ll Q$ and both sides are $+\infty$. Suppose $\alpha \ll \beta$ and let $h = d\alpha/d\beta$. If the conditional absolute continuity $\nu(p) \ll \mu(p)$ fails on a set of positive α-measure, then again both sides are $+\infty$. Otherwise,

$$\frac{dY}{dQ}(p, j) = h(p)\frac{\nu_j(p)}{\mu_j(p)} \qquad Y\text{-a.e.}$$

The two logarithmic terms can be integrated separately without an $\infty - \infty$ ambiguity. Indeed,

$$\int_{\mathcal{U}} (\log h)_-\, d\alpha = \int_{\{h<1\}} h(-\log h)\, d\beta \le \frac{1}{e},$$

since $t(-\log t) \le 1/e$ for $0 < t \le 1$. Likewise, for every p,

$$\sum_j \nu_j(p) \left(\log \frac{\nu_j(p)}{\mu_j(p)}\right)_- \le \frac{1}{e},$$

because on coordinates with $0 < \nu_j < \mu_j$, writing $t_j = \nu_j/\mu_j$ gives $\nu_j(-\log t_j) = \mu_j t_j(-\log t_j) \le \mu_j/e$. Thus both negative parts are integrable.

Consequently,

$$\begin{aligned} D_{\mathrm{KL}}(Y\|Q) &= \int_{\mathcal{U}} \log h(p)\, d\alpha(p) + \int_{\mathcal{U}} \sum_j \nu_j(p) \log \frac{\nu_j(p)}{\mu_j(p)}\, d\alpha(p) \\ &= D_{\mathrm{KL}}(\alpha\|\beta) + \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu(p)\|\mu(p)\big)\, d\alpha(p), \end{aligned}$$

which proves equation 37 in the extended nonnegative reals. □

C.2 EXACT JOINT PROXIMAL COMPLETION

Lemma C.2 (Exact joint proximal completion)**.** *Let $z^\circ = (\pi^\circ, Q^\circ)$ with $Q^\circ(dp, j) = \alpha^\circ(dp)\nu_j^\circ(p)$. Let $s \in \mathbb{R}^m$, let $c_j : \mathcal{U} \to \mathbb{R}$ be Borel with $c_{\min} \le c_j(p) \le c_{\max}$, and let $\eta > 0$ be finite. Set $a := 1 + \eta\tau$ and define*

$$w_i := (\pi_i^\circ)^{1/a} r_i^{\eta\tau/a} e^{\eta s_i/a}, \qquad W := \sum_i w_i, \qquad \pi_i^+ := w_i/W,$$

$$b_j(p) := (\nu_j^\circ(p))^{1/a} r_j^{\eta\tau/a} e^{-\eta c_j(p)/a}, \qquad Z(p) := \sum_j b_j(p), \qquad \nu_j^+(p) := b_j(p)/Z(p).$$

With

$$C := \int_{\mathcal{U}} Z(p)^a\, d\alpha^\circ(p), \qquad \alpha^+(dp) := \frac{Z(p)^a}{C}\,\alpha^\circ(dp), \qquad Q^+(dp, j) := \alpha^+(dp)\nu_j^+(p),$$

all normalizers are positive and finite, and $z^+ := (\pi^+, Q^+)$ is the unique minimizer of

$$\mathcal{J}(u) := D_{z^\circ}(u) + \eta\langle(-s, c), u\rangle + \eta H(u), \qquad u = (x, Y) \in \Delta_m \times \mathcal{Q}.$$

Moreover, $D_{z^\circ}(z^+) < \infty$, $Q^+ \sim Q^\circ$, and π^+ has the same support as π°. For every $u = (x, Y)$ with $D_{z^\circ}(u) < \infty$,

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(z^+) = {} & aD_{\mathrm{KL}}(x\|\pi^+) + D_{\mathrm{KL}}(\alpha_Y\|\alpha^+) \\ & + a\int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Y(p)\|\nu^+(p)\big)\, d\alpha_Y(p). \end{aligned} \tag{38}$$

Proof. Well-defined update. The weights are Borel, and their positive coordinates coincide with those of the center. Hölder's inequality and $\sum_j (\nu_j^\circ(p))^{1/a} \geq 1$ give

$$e^{-\eta c_{\max}/a} r_{\min}^{\eta\tau/a} \leq Z(p) \leq e^{-\eta c_{\min}/a}. \tag{39}$$

Moreover,

$$b_j(p) > 0 \quad \Longleftrightarrow \quad \nu_j^\circ(p) > 0,$$

because $r_j > 0$ and the exponential factor is positive. Hence $\nu_j^+(p) > 0$ if and only if $\nu_j^\circ(p) > 0$. Together with $\alpha^+ \sim \alpha^\circ$, this implies $Q^+ \sim Q^\circ$.

The conditional formula and equation 39 imply $\nu_j^+(p) \leq M_0(\nu_j^\circ(p))^{1/a}$ for a finite constant M_0 independent of p, j. Consequently,

$$\sum_{j:\, \nu_j^\circ(p)>0} \nu_j^+(p) |\log \nu_j^\circ(p)| \leq M_0 m \sup_{0<v\leq 1} v^{1/a} |\log v| = \frac{M_0 m a}{e}.$$

On the common support of $\nu^+(p)$ and $\nu^\circ(p)$, the update formula gives

$$\log \frac{\nu_j^+(p)}{\nu_j^\circ(p)} = -\frac{a-1}{a} \log \nu_j^\circ(p) + \frac{\eta\tau}{a} \log r_j - \frac{\eta}{a} c_j(p) - \log Z(p).$$

Hence

$$D_{\mathrm{KL}}\big(\nu^+(p) \| \nu^\circ(p)\big) = \sum_{j:\, \nu_j^\circ(p)>0} \nu_j^+(p) \log \frac{\nu_j^+(p)}{\nu_j^\circ(p)}.$$

The term involving $\log \nu_j^\circ(p)$ is uniformly integrable by the preceding bound, while the remaining terms are uniformly bounded because r has full support, c is bounded, and equation 39 bounds $Z(p)$ above and away from zero. Therefore

$$\sup_{p\in\mathcal{U}} D_{\mathrm{KL}}\big(\nu^+(p) \| \nu^\circ(p)\big) < \infty.$$

The chain rule in Lemma C.1 gives $D_{\mathrm{KL}}(Q^+ \| Q^\circ) < \infty$; the leader KL is finite on its support face. Hence $D_{z^\circ}(z^+) < \infty$.

Completion and uniqueness. For $D_{z^\circ}(u) < \infty$, apply Lemma C.1 and substitute

$$\log \nu_j^+ = \frac{\log \nu_j^\circ + \eta\tau \log r_j - \eta c_j}{a} - \log Z, \qquad \log \frac{d\alpha^+}{d\alpha^\circ} = a \log Z - \log C$$

on the supported coordinates, together with the leader formula. By the KL chain rule,

$$D_{z^\circ}(u) = D_{\mathrm{KL}}(x\|\pi^\circ) + D_{\mathrm{KL}}(\alpha_Y\|\alpha^\circ) + \int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Y(p)\|\nu^\circ(p)\big)\, d\alpha_Y(p).$$

Hence

$$\begin{aligned}\mathcal{J}(u) = &\Big[D_{\mathrm{KL}}(x\|\pi^\circ) + \eta\tau D_{\mathrm{KL}}(x\|r) - \eta s^\top x\Big] \\ &+ D_{\mathrm{KL}}(\alpha_Y\|\alpha^\circ) \\ &+ \int_{\mathcal{U}} \Big[D_{\mathrm{KL}}(\nu_Y(p)\|\nu^\circ(p)) + \eta\tau D_{\mathrm{KL}}(\nu_Y(p)\|r) + \eta c(p)^\top \nu_Y(p)\Big]\, d\alpha_Y(p).\end{aligned}$$

The leader completion gives

$$D_{\mathrm{KL}}(x\|\pi^\circ) + \eta\tau D_{\mathrm{KL}}(x\|r) - \eta s^\top x = a D_{\mathrm{KL}}(x\|\pi^+) - a \log W.$$

For α_Y-almost every p,

$$D_{\mathrm{KL}}(\nu_Y(p)\|\nu^\circ(p)) + \eta\tau D_{\mathrm{KL}}(\nu_Y(p)\|r) + \eta c(p)^\top \nu_Y(p) = a D_{\mathrm{KL}}(\nu_Y(p)\|\nu^+(p)) - a \log Z(p).$$

Since

$$\log \frac{d\alpha^+}{d\alpha^\circ} = a \log Z - \log C,$$

we have

$$D_{\mathrm{KL}}(\alpha_Y\|\alpha^\circ) - a\int_{\mathcal{U}} \log Z\, d\alpha_Y = D_{\mathrm{KL}}(\alpha_Y\|\alpha^+) - \log C.$$

Combining these terms gives

$$\begin{aligned}\mathcal{J}(u) = & -a\log W - \log C + aD_{\mathrm{KL}}(x\|\pi^+) + D_{\mathrm{KL}}(\alpha_Y\|\alpha^+) \\ & + a\int_{\mathcal{U}} D_{\mathrm{KL}}\big(\nu_Y(p)\|\nu^+(p)\big)\, d\alpha_Y(p).\end{aligned}$$

At $u = z^+$ the KL terms vanish, so $\mathcal{J}(z^+) = -a\log W - \log C$ and equation 38 follows. The nonnegative terms vanish simultaneously only when $x = \pi^+$ and $Y = Q^+$. Comparators with $D_{z^\circ}(u) = \infty$ have $\mathcal{J}(u) = \infty$, since the direction pairing and H are bounded. Thus z^+ is the unique minimizer over the entire feasible domain. □

C.3 Same-center Mirror-Prox estimate

Recall $F(\pi, Q) = (-s(Q), c_\pi)$, where $c_\pi(p) = A(p)^\top\pi$. For a finite signed joint measure M, extend the score map linearly by

$$s_i(M) := \sum_{j=1}^m \int_{\mathcal{U}} A_{ij}(p)\, M(dp, j).$$

For a leader increment h, likewise write

$$c_h(p) := A(p)^\top h.$$

For a leader increment h, a signed joint measure M, and a bounded direction (v, c), define

$$N(h, M)^2 = \|h\|_1^2 + \|M\|_{\mathrm{var}}^2, \qquad N_*(v, c)^2 = \|v\|_\infty^2 + \|c\|_\infty^2,$$

where $\|M\|_{\mathrm{var}}$ is the total variation norm and $\|c\|_\infty = \sup_{p,j}|c_j(p)|$. By Hölder's inequality for the leader block and the duality between bounded functions and total variation for the joint block,

$$\begin{aligned}|\langle(v, c), (h, M)\rangle| &\le \|v\|_\infty\|h\|_1 + \|c\|_\infty\|M\|_{\mathrm{var}} \\ &\le N_*(v, c)\, N(h, M),\end{aligned}$$

where the last step is Cauchy–Schwarz in $\mathbb{R}^2$.

Lemma C.3 (Same-center Mirror-Prox estimate)**.** *For bounded kernels, let $L_F \ge 0$ satisfy*

$$N_*(F(z) - F(w)) \le L_F N(z - w) \quad \textit{for all feasible } z, w. \tag{40}$$

Given a center z, form the predictor and corrector

$$\begin{aligned}\widetilde{z} &= \arg\min_{w\in\Delta_m\times\mathcal{Q}} \big\{D_z(w) + \eta\langle F(z), w\rangle + \eta H(w)\big\}, \\ z^+ &= \arg\min_{w\in\Delta_m\times\mathcal{Q}} \big\{D_z(w) + \eta\langle F(\widetilde{z}), w\rangle + \eta H(w)\big\}.\end{aligned}$$

If $\eta > 0$ is finite and $\eta L_F \le 1$, then every $u = (x, Y)$ with $D_z(u) < \infty$ satisfies

$$\eta\big[L(x, \widetilde{Q}) - L(\widetilde{\pi}, Y)\big] \le D_z(u) - D_{z^+}(u). \tag{41}$$

For unit-range kernels, $L_F = 1$ is valid.

Proof. Proximal inequality. For a proximal output w from center z with bounded direction ξ, define

$$\mathcal{J}_\xi(u') := D_z(u') + \eta\langle\xi, u'\rangle + \eta H(u').$$

Lemma C.2, Lemma C.1, and $1 + \eta\tau \ge 1$ give

$$\mathcal{J}_\xi(u') - \mathcal{J}_\xi(w) \ge D_w(u') \qquad \text{whenever } D_z(u') < \infty.$$

The left-hand side is finite, so $D_w(u') < \infty$. Rearranging yields

$$\eta\big[\langle\xi, w - u'\rangle + H(w) - H(u')\big] \le D_z(u') - D_w(u') - D_z(w). \tag{42}$$

Same-center cancellation. Apply equation 42 to the predictor with comparison point z^+, and to the corrector with comparison point u. Adding the two inequalities cancels $D_z(z^+)$ and $H(z^+)$. For the remaining direction terms,

$$\begin{aligned}&\langle F(z), \widetilde{z} - z^+\rangle + \langle F(\widetilde{z}), z^+ - u\rangle \\ &\quad = \langle F(\widetilde{z}), \widetilde{z} - u\rangle - \langle F(\widetilde{z}) - F(z), \widetilde{z} - z^+\rangle.\end{aligned}$$

Rearranging therefore gives

$$\begin{aligned}&\eta\big[\langle F(\widetilde{z}), \widetilde{z} - u\rangle + H(\widetilde{z}) - H(u)\big] \\ &\quad \le D_z(u) - D_{z^+}(u) - D_{\widetilde{z}}(z^+) - D_z(\widetilde{z}) \\ &\qquad + \eta\langle F(\widetilde{z}) - F(z), \widetilde{z} - z^+\rangle. \end{aligned} \tag{43}$$

Cross-term bound. For unit-range kernels, direct integration gives

$$\|s(M)\|_\infty \le \|M\|_{\mathrm{var}}, \qquad \sup_{p,j} |[A(p)^\top h]_j| \le \|h\|_1.$$

Squaring and adding verifies equation 40 with $L_F = 1$. Pinsker's inequality, applied to both blocks, gives

$$D_z(w) \ge \frac{1}{2} N(w - z)^2. \tag{44}$$

Set $\delta = N(\widetilde{z} - z)$ and $\delta_+ = N(\widetilde{z} - z^+)$. The pairing bound, equation 40, and equation 44 imply

$$\begin{aligned}&\eta\langle F(\widetilde{z}) - F(z), \widetilde{z} - z^+\rangle - D_{\widetilde{z}}(z^+) - D_z(\widetilde{z}) \\ &\quad \le \eta L_F \delta\delta_+ - \frac{1}{2}(\delta^2 + \delta_+^2) \le 0,\end{aligned} \tag{45}$$

where the last inequality uses $\eta L_F \le 1$.

Finally, let $z = (\pi, Q)$ and $u = (x, Y)$. Using

$$F(\pi, Q) = (-s(Q), c_\pi), \qquad \langle c_\pi, Q\rangle = \pi^\top s(Q), \qquad \langle c_\pi, Y\rangle = \pi^\top s(Y),$$

we obtain

$$\begin{aligned}\langle F(z), z - u\rangle &= -s(Q)^\top(\pi - x) + \langle c_\pi, Q - Y\rangle \\ &= x^\top s(Q) - \pi^\top s(Y).\end{aligned}$$

Therefore,

$$\begin{aligned}&\langle F(z), z - u\rangle + H(z) - H(u) \\ &\quad = x^\top s(Q) - \pi^\top s(Y) + \tau D_{\mathrm{KL}}(\pi\|r) - \tau D_{\mathrm{KL}}(x\|r) + \tau\Psi(Q) - \tau\Psi(Y) \\ &\quad = L(x, Q) - L(\pi, Y).\end{aligned} \tag{46}$$

□

D Additional Experimental Details

D.1 Finite preference uncertainty on PersonalLLM

We evaluate finite-scenario preference uncertainty on PersonalLLM (Zollo et al., 2025), where multiple released reward models provide distinct preference scenarios over the same candidate responses. This experiment tests both optimization of the finite robust game and the tradeoff between nominal and worst-case performance as the assumed uncertainty level changes.

D.1.1 Data construction and preference kernels

PersonalLLM contains $9{,}402$ training rows, each with eight candidate responses scored by ten released reward-model providers. We remove 33 prompts overlapping the official test split and 193 additional duplicate normalized prompts, leaving $9{,}176$ eligible rows. We reserve 256 prompts for score calibration, 256 for development, and $8{,}000$ for evaluation; the remaining 664 are unused. All

Table 2: Reward-model providers used in the PersonalLLM experiment.

Family / source label	Released provider identifiers
Gemma	`gemma_2b`, `gemma_7b`
Mistral	`mistral_raft`, `mistral_ray`, `mistral_weqweasdas`
Llama-3	`llama3_sfairx`
OASST / DeBERTa	`oasst_deberta_v3`
Beaver	`beaver_7b`
OASST / Pythia	`oasst_pythia_7b`, `oasst_pythia_1b`

evaluation prompts are distinct after normalization and have no normalized-prompt overlap with the official test split.

For each provider k, we estimate an RMS scale $\widehat{\sigma}_k$ from the $256\binom{8}{2}$ pairwise score differences in the calibration set and normalize its scores accordingly. All released score columns are oriented so that larger values are preferred; the released Beaver wrapper already negates its original cost score, so no additional sign reversal is applied. Writing $\widetilde{s}^k_{x,i}$ for the calibrated score, we construct

$$(A^k_x)_{ij} = \sigma\big(\widetilde{s}^k_{x,i} - \widetilde{s}^k_{x,j}\big), \qquad (A^k_x)_{ji} = 1 - (A^k_x)_{ij}, \qquad (A^k_x)_{ii} = \tfrac{1}{2}.$$

Calibration uses only the held-out calibration cohort. The nominal kernel is the probability-kernel mean $A_{N,x} = K^{-1}\sum_k A^k_x$, with $K = 10$.

Let $\Delta^k_x := A^k_x - A_{N,x}$. For $\rho \in [0,1]$, we consider

$$\mathcal{U}_x(\rho) := A_{N,x} + \rho\ \mathrm{conv}\{\Delta^1_x, \dots, \Delta^K_x\}.$$

Thus $\rho = 0$ gives the nominal game, while $\rho = 1$ gives the full convex hull of the ten provider kernels. Because $J(\pi, \nu; A)$ is affine in A, the worst case over $\mathcal{U}_x(\rho)$ can be attained at one of the generating kernels $A_{N,x} + \rho\Delta^k_x$. The nominal kernel belongs to every radius because $\sum_k \Delta^k_x = 0$.

D.1.2 Optimization and evaluation protocol

Each prompt defines an independent eight-action game; no parameters are shared across prompts. EJMP uses the same-old-center predictor–corrector update from Algorithm 1. We average predictor leaders arithmetically and average the adversary directly in the joint q-representation. Nash-MD optimizes only $A_{N,x}$ and returns its prescribed last iterate. All methods use the uniform reference $r = (1/8, \dots, 1/8)$ and $\tau = 0.05$.

For fixed π_x, provider k, and evaluation radius ρ, define $A_{x,k}(\rho) := A_{N,x} + \rho\Delta^k_x$. The inner follower problem is available in closed form:

$$h_{xk,\rho}(\pi_x) = -\tau D_{\mathrm{KL}}(\pi_x \| r) - \tau \log \sum_{j=1}^{8} r_j \exp\left(-\frac{[\pi_x^\top A_{x,k}(\rho)]_j}{\tau}\right).$$

Hence evaluation requires no iterative inner best response. Our primary metric is the prompt-averaged robust value

$$\bar{g}_\rho(\pi) := \frac{1}{N}\sum_{x=1}^{N} \min_{k \in [K]} h_{xk,\rho}(\pi_x),$$

which averages each prompt's worst-case regularized game value. Final policies are evaluated without further optimization over $\rho_{\mathrm{eval}} \in \{0, 0.1, \dots, 1\}$, with 0.25 and 0.75 additionally included for matched-radius comparisons.

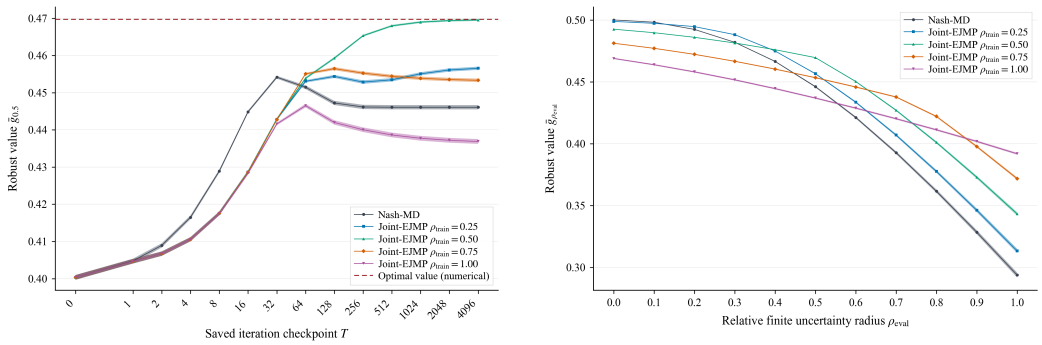
Confidence intervals use $10{,}000$ paired prompt-bootstrap resamples. Each replicate uses the same prompt indices across methods and radii. We use NumPy PCG64 with seed 3505102 and report pointwise 95% percentile intervals.

D.1.3 Results

Figure 1 summarizes the two main empirical findings. At the common evaluation radius $\rho_{\mathrm{eval}} = 0.5$, EJMP trained at the matching radius approaches the independently computed numerical optimum,

Table 3: PersonalLLM optimization and evaluation settings.

Setting	Value
Evaluation prompts	8,000
Candidates / providers	$m = 8,\ K = 10$
Reference r	Uniform on 8 responses
Bradley–Terry scale	$\beta = 1$
KL temperature	$\tau = 0.05$
EJMP step size	$\eta = 0.5$
Nash-MD parameter	$c = 4$
Optimization horizon	$T = 4096$
Training radii	$\{0.25, 0.50, 0.75, 1.00\}$
Evaluation radii	$\{0, 0.1, \ldots, 1\} \cup \{0.25, 0.75\}$
Bootstrap resamples	10,000



(a) Optimization at $\rho_{\text{eval}} = 0.5$.

(b) Final-policy radius sweep.

Figure 1: PersonalLLM robust values. Left: saved EJMP checkpoints under a common evaluation radius; the dashed line denotes the independently computed numerical optimum. Right: final policies evaluated across uncertainty radii without reoptimization. Shading denotes pointwise 95% paired-bootstrap intervals.

while both smaller and larger training radii attain lower final robust values (Figure 1a). Thus, enlarging the training uncertainty set is not uniformly beneficial under a fixed evaluation criterion.

The radius sweep in Figure 1b shows the corresponding nominal–robustness tradeoff. At each tested matched radius $\rho_{\text{eval}} \in \{0.25, 0.50, 0.75, 1.00\}$, the policy trained at the same radius achieves the largest robust value among the four EJMP variants. The matched gain over Nash-MD increases from 0.0049 at $\rho_{\text{eval}} = 0.25$ to 0.0980 at full provider uncertainty (Table 4). Conversely, the policy trained at $\rho_{\text{train}} = 1$ loses 0.0311 relative to Nash-MD under the nominal kernel. The experiment therefore exhibits a clear tradeoff between nominal performance and protection against stronger preference variation.

Figure 2 isolates the paired improvement over Nash-MD. Policies trained with larger radii are disadvantaged near the nominal kernel but become favorable as evaluation uncertainty increases, with the crossing region shifting outward as the training radius grows.

To validate the optimization reference in Figure 1a, we independently solve each promptwise robust problem at $\rho_{\text{eval}} = 0.5$:

$$\max_{\pi \in \Delta_8,\, t} t \qquad \text{s.t.} \qquad t \le h_{xk,0.5}(\pi) \quad \forall k \in [10].$$

We use SLSQP with analytic gradients and bound each optimum using feasible primal and joint-adversarial witnesses. All 8,000 problems satisfy the predeclared 10^{-8} primal–dual tolerance; the largest observed promptwise gap is 8.04×10^{-13}. We additionally cross-check 16 problems with CVXPY/CLARABEL and 16 selected witness pairs using 50-digit arithmetic. These checks support the numerical accuracy of the dashed reference for the fixed empirical problem, rather than providing a formal interval-arithmetic certificate.

Table 4: Prompt-averaged robust value at matched training and evaluation radii. Higher is better.

ρ_{eval}	0.25	0.50	0.75	1.00
Nash-MD	0.4878	0.4461	0.3775	0.2939
EJMP	**0.4927**	**0.4696**	**0.4333**	**0.3919**
Gain over Nash-MD	+0.0049	+0.0235	+0.0558	+0.0980

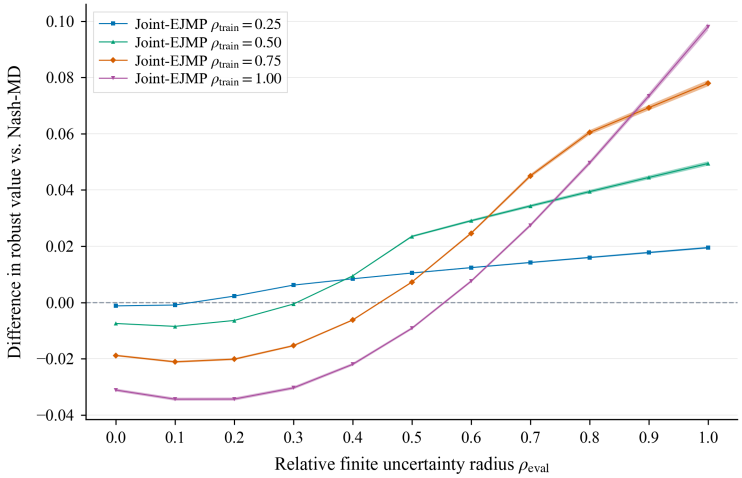


Figure 2: EJMP minus Nash-MD robust value. Positive values favor EJMP; shading denotes pointwise 95% paired-bootstrap intervals.

Finally, Table 5 reports late-stage optimization diagnostics at matched radii. The change in robust value between $T = 2048$ and 4096 is below 2.1×10^{-4} in every case, while the computed primal–dual gaps remain below 8×10^{-4}.

D.2 Continuous preference uncertainty on IMDb

We provide the implementation and evaluation details omitted from Section 6, including candidate generation, reference construction, stress-test distributions, and additional horizon and uniform-reference results.

D.2.1 Cohort and candidate generation

We construct a fixed best-of-m continuation benchmark from IMDb (Maas et al., 2011). The first 1,000 source rows form the development region; subsequent rows are scanned in deterministic order until 4,000 valid records are accepted. Each prompt has exactly $m = 8$ normalized-distinct continuations generated by FLAN-T5-Large (Chung et al., 2024). Eight fixed instructions cover analytical, concise, balanced, and neutral styles; temperatures cycle over $\{0.7, 0.9, 1.1, 1.3\}$, with top-$p = 0.95$ and at most 150 new tokens. Generation is seeded by source-row index, and rows failing to produce eight valid continuations are rejected rather than padded.

The final cohort requires 424 rejections and contains 3,993 distinct prompt texts. It is disjoint from the accepted 200-prompt development cache by source-row identity, although no additional text-level deduplication is performed. Because the selected source rows lie in IMDb's negative-label block, we treat this experiment as a controlled continuation benchmark rather than a sentiment-representative sample.

Table 5: Late-stage EJMP diagnostics at matched radii.

ρ_{train}	$\bar{g}_{2048}$	$\bar{g}_{4096}$	$\Delta\bar{g}$	Gap
0.25	0.492516	0.492725	2.09×10^{-4}	0.000648
0.50	0.469408	0.469580	1.71×10^{-4}	0.000682
0.75	0.433148	0.433329	1.82×10^{-4}	0.000718
1.00	0.391677	0.391876	1.99×10^{-4}	0.000754

Table 6: **IMDb experimental configuration.**

Quantity	Value
Formal records	4,000
Candidates per prompt	$m = 8$
Generator	FLAN-T5-Large
Nominal group mixture	$(0.5, 0.3, 0.2)$
Bradley–Terry scale	$\beta = 5$
Regularization	$\tau = 0.01$
Training radii	$\{.005, .01, .02, .05, .10, .15\}$
Evaluation radii	$\{.05, .10, .20, .30, .50\}$
EJMP	$\eta = 0.5,\ T \in \{256, 512, 1024, 2048\}$
EJMP quadrature	48×6
Nash-MD	$T = 200,\ \eta = 0.30$
Mixture-KL BR	$T = 300,\ 8$ inner steps
Control step sizes	$\eta_\pi = 0.25,\ \eta_\gamma = 0.15$
Dual range	$[0.02, 500]$
Evaluation kernels	$M = 600$ per prompt/radius
Bootstrap	2,000 prompt resamples

D.2.2 Preference kernels and references

We score each candidate under three deterministic profiles: comprehensive/analytical, efficient/concise, and balanced/nuanced. For profile $g \in \{A, B, C\}$, scalar scores $s_i^{(g)}$ induce

$$p_{ij}^{(g)} = \sigma\Big(\beta(s_i^{(g)} - s_j^{(g)})\Big), \qquad \beta = 5.$$

With nominal weights $w_0 = (0.5, 0.3, 0.2)$, $A(w_0) = 0.5A^A + 0.3A^B + 0.2A^C$, and robust training uses

$$\mathcal{U}_\rho = \{w \in \Delta_3 : D_{\text{KL}}(w\|w_0) \le \rho\}.$$

Our primary reference is prompt-specific and derived from the frozen generator's candidate log likelihoods:

$$\widetilde{r}_i = \frac{\exp(\ell_i/T_{\text{ref}})}{\sum_j \exp(\ell_j/T_{\text{ref}})}, \qquad T_{\text{ref}} = 20.$$

We apply a 10^{-6} floor and renormalize. We also evaluate a uniform reference $r_i = 1/8$. Within each comparison, the robust policy and Nash-MD use the same reference.

D.2.3 Optimization

Table 6 summarizes the experimental configuration. EJMP uses the same-old-center predictor–corrector recurrence from Section 5, fixed quadrature, and arithmetic averaging of predictor joint laws. Nash-MD uses optimistic entropic updates and returns its second-half leader average, while the mixture-KL best-response control alternates outer leader/dual updates with inner follower/scenario updates. All four prespecified EJMP horizons are retained, with $T = 2048$ used for the main-text endpoint results.

D.2.4 Stress-test distributions and evaluation

We evaluate on two stress-test distributions that are deliberately different from the mixture-KL ambiguity set used for training. For each prompt x, let $P_x^\star := A_x(w_0)$ denote the nominal preference

kernel, and let $p_x^\star \in (0,1)^{\binom{m}{2}}$ collect its upper-triangular entries. For reciprocal kernels P and $P_x^\star$, define

$$\mathcal{D}_{\mathrm{B}}(P\|P_x^\star) := \sum_{1\le i<j\le m} D_{\mathrm{KL}}^{\mathrm{Bern}}\big(P_{ij}\|P_{x,ij}^\star\big)\,, \qquad D_{\mathrm{KL}}^{\mathrm{Bern}}(q\|p) := q\log\frac{q}{p}+(1-q)\log\frac{1-q}{1-p}. \tag{47}$$

Since $m = 8$, there are 28 independent upper-triangular pairwise probabilities.

E1: pairwise-logit KL perturbations. For each evaluation draw, sample $z \sim \mathcal{N}(0, I_{28})$, set $d = z/\|z\|_2$, and define

$$p_x(\lambda) = \sigma(\mathrm{logit}(p_x^\star) + \lambda d)\,, \qquad \lambda \ge 0.$$

We choose λ by bisection so that

$$\mathcal{D}_{\mathrm{B}}(P_x(\lambda)\|P_x^\star) = \rho_{\mathrm{eval}}, \tag{48}$$

where $P_x(\lambda)$ is completed by reciprocity, $P_{x,ji}(\lambda) = 1 - P_{x,ij}(\lambda)$, with diagonal entries $1/2$. Thus E1 samples random logit-space rays whose endpoints lie on the prescribed Bernoulli–KL boundary, rather than uniformly sampling that boundary.

E2: annotator-mixture shifts. For each proposal, sample

$$\widetilde{w} \sim \mathrm{Dirichlet}(\kappa w_0 + 0.05\,\mathbf{1}_3)\,, \qquad \kappa = 2,$$

where $\mathbf{1}_3 = (1,1,1)$, and form

$$P_x(\widetilde{w}) := \sum_{g=1}^{3} \widetilde{w}_g A_x^g.$$

We retain the first proposal satisfying

$$\frac{1}{2}\rho_{\mathrm{eval}} \le D_B(P_x(\widetilde{w})\|P_x^\star) \le \rho_{\mathrm{eval}}.$$

Up to 200 proposals are attempted. If none falls in this outer portion of the ball, the sampler returns the first feasible proposal satisfying $D_B(P_x(\widetilde{w})\|P_x^\star) \le \rho_{\mathrm{eval}}$; if no feasible proposal is found, it returns the nominal mixture $\widetilde{w} = w_0$. Thus E2 perturbs evaluator composition while controlling the resulting kernel shift through the same Bernoulli–KL measure.

For every retained evaluation kernel $P_{x,\ell}$, let $\pi_{N,x}$ denote the matched-reference Nash-MD policy and define

$$W_{x,\ell} := \pi_x^\top P_{x,\ell}\pi_{N,x}.$$

Our primary evaluation statistic is

$$\mathrm{WinFrac} := \frac{1}{NM}\sum_{x=1}^{N}\sum_{\ell=1}^{M}\mathbf{1}\left\{W_{x,\ell} > \frac{1}{2}\right\}, \tag{49}$$

the fraction of sampled prompt–kernel pairs for which EJMP has a positive pairwise preference advantage over Nash-MD. We also report $\mathrm{MeanWin} := (NM)^{-1}\sum_{x=1}^{N}\sum_{\ell=1}^{M} W_{x,\ell}$, the average head-to-head preference probability. All methods use the same fixed evaluation banks. Confidence intervals use paired prompt-bootstrap resampling, with the evaluation-bank and bootstrap settings reported in Table 6.

D.2.5 Additional results

Table 7 summarizes the effect of optimization horizon over the complete 6×5 radius grid. The SFT-reference results improve sharply between $T = 256$ and 512 and continue improving thereafter; the uniform-reference results show an even larger early-horizon effect.

Figure 3 shows the same horizon effect cell-by-cell for the SFT-derived reference. The improvement from $T = 256$ to 512 occurs broadly across the grid, while annotator-mixture shifts remain more sensitive to larger evaluation radii.

Table 7: **EJMP horizon dependence.** Entries average all 30 radius pairs.

Reference	T	KL perturbations		Annotator-mixture shifts	
		WinFrac	MeanWin	WinFrac	MeanWin
SFT	256	.766	.512	.736	.514
SFT	512	.956	.535	.926	.538
SFT	1024	.972	.547	.944	.550
SFT	2048	.974	.552	.948	.556
Uniform	256	.413	.499	.394	.499
Uniform	512	.801	.505	.781	.505
Uniform	1024	.822	.508	.803	.508
Uniform	2048	.825	.509	.808	.510

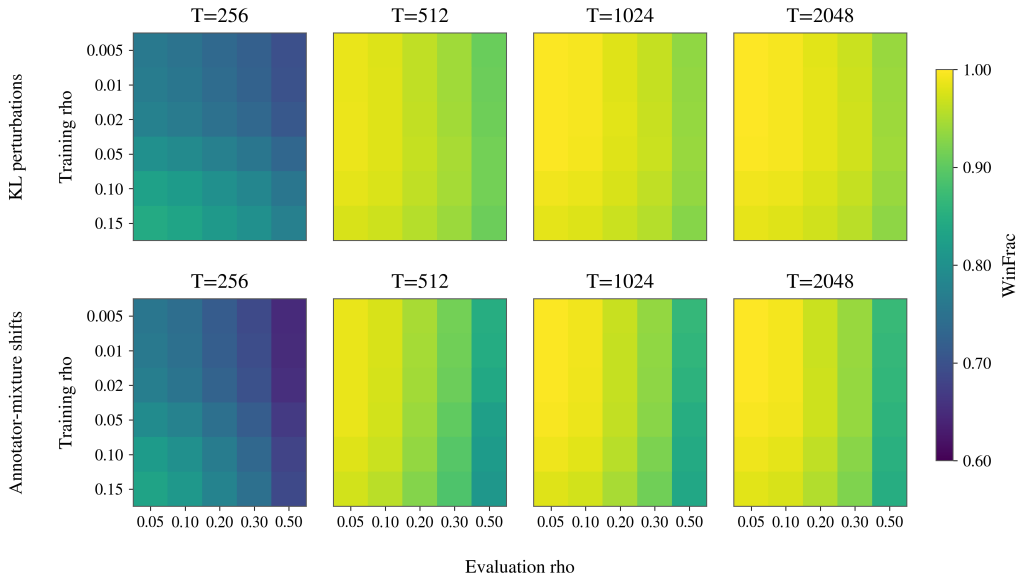


Figure 3: SFT-reference IMDb WinFrac across optimization horizons. Columns show $T \in \{256, 512, 1024, 2048\}$; rows show KL perturbations (top) and annotator-mixture shifts (bottom).

Table 8 gives the final uniform-reference grid. Performance is strong at moderate training radii but deteriorates at $\rho_{\text{train}} = .15$, showing that a larger training ambiguity set need not improve head-to-head preference performance.

Finally, Table 9 illustrates why both statistics are reported. With the SFT-derived reference, EJMP improves both WinFrac and MeanWin over the mixture-KL best-response control. With the uniform reference, the control attains the higher WinFrac, whereas EJMP attains the higher MeanWin.

At $T = 2048$, the paired SFT-reference EJMP-minus-control differences are $+0.185/+0.046$ in WinFrac/MeanWin under KL perturbations and $+0.180/+0.050$ under annotator-mixture shifts; all four paired 95% bootstrap intervals exclude zero. Under the uniform reference, EJMP has lower WinFrac ($-0.107/-0.087$) but higher MeanWin ($+0.0054/+0.0057$) for the two stress tests.

Table 8: **Uniform-reference EJMP at $T = 2048$.** Entries are WinFrac against matched-reference Nash-MD.

	KL perturbations					Annotator-mixture shifts				
$\rho_{\text{train}} \backslash \rho_{\text{eval}}$	.05	.10	.20	.30	.50	.05	.10	.20	.30	.50
.005	.997	.983	.948	.917	.869	.987	.956	.906	.868	.816
.010	.997	.984	.950	.920	.872	.989	.959	.909	.871	.819
.020	.998	.987	.956	.926	.879	.992	.967	.918	.879	.826
.050	.995	.984	.955	.928	.883	.990	.973	.942	.911	.858
.100	.756	.747	.727	.711	.685	.731	.728	.737	.733	.709
.150	.416	.427	.440	.448	.458	.406	.418	.457	.485	.501

Table 9: **Method ordering under WinFrac and MeanWin.** Values average the full 30-cell grid.

Reference	Method	KL perturbations		Annotator-mixture shifts	
		WinFrac	MeanWin	WinFrac	MeanWin
SFT	Mixture-KL BR	.789	.506	.768	.506
SFT	EJMP	**.974**	**.552**	**.948**	**.556**
Uniform	Mixture-KL BR	**.931**	.503	**.895**	.504
Uniform	EJMP	.825	**.509**	.808	**.510**