Beyond RLHF and NLHF: Population-Proportional Alignment under an Axiomatic Framework

Kihyun Kim[1*] **Jiawei Zhang**[1,2*†] **Asuman Ozdaglar**[1] **Pablo A. Parrilo**[1]
[1] MIT LIDS [2] University of Wisconsin–Madison
{kihyun, jwzhang, asuman, parrilo}@mit.edu

Abstract

Conventional preference learning methods often prioritize opinions held more widely when aggregating preferences from multiple evaluators. This may result in policies that are biased in favor of some types of opinions or groups and susceptible to strategic manipulation. To address this issue, we develop a novel preference learning framework capable of aligning aggregate opinions and policies proportionally with the true population distribution of evaluator preferences. Grounded in social choice theory, our approach infers the feasible set of evaluator population distributions directly from pairwise comparison data. Using these estimates, the algorithm constructs a policy that satisfies foundational axioms from social choice theory, namely monotonicity and Pareto efficiency, as well as our newly-introduced axioms of population-proportional alignment and population-bounded manipulability. Moreover, we propose a soft-max relaxation method that smoothly trade-offs population-proportional alignment with the selection of the Condorcet winner (which beats all other options in pairwise comparisons). Finally, we validate the effectiveness and scalability of our approach through experiments on both tabular recommendation tasks and large language model alignment.

1 Introduction

Aligning artificial intelligence (AI) systems with complex human preferences is a growing priority in fields such as robotics [1, 2], recommendation systems [3], and large language models (LLMs) [4–6]. A key challenge in this endeavor is how to infer and represent such preferences accurately, particularly when they are only available through incomplete signals like pairwise comparisons. This has prompted reinforcement learning from human feedback (RLHF), which has become a widely used framework for preference learning [6, 7]. RLHF streamlines the alignment process by first learning a reward model that assigns scalar scores to different alternatives, typically trained using maximum likelihood estimation under the Bradley–Terry (BT) model. In the second stage, a policy is optimized through reinforcement learning to maximize the expected rewards, guiding the system toward behaviors aligned with human preferences.

Despite its practical success and simplicity, the standard RLHF framework rests on a critical assumption that complex human preferences can be captured by a single scalar reward. Recent research highlights that this assumption often breaks down, especially when human feedback reflects inconsistent or conflicting judgments across evaluators [8]. In particular, RLHF struggles in scenarios involving intransitive or cyclic preferences, where no clear ranking among alternatives can be established, leading to failures in accurately modeling the underlying preferences [9, 10]. To address these

[*] Equal contribution.
[†] Corresponding author.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

limitations, a game-theoretic framework called Nash learning from human feedback (NLHF) has been introduced [9–12]. NLHF reframes preference learning as a two-player constant-sum game and identifies equilibrium policies that no competing policy can outperform, regardless of the complexity of the underlying preferences.

Nevertheless, both RLHF and NLHF frameworks remain limited in their ability to address another critical issue: the proportional alignment of evaluator preferences. When preferences are aggregated across multiple evaluator groups with distinct viewpoints, both RLHF and NLHF tend to yield policies that do not adequately reflect the full distribution of the evaluator population [8]. To address these challenges, recent research has turned to social choice theory-oriented approaches, such as maximizing the minimum satisfaction across evaluator groups [8, 13] and optimizing social welfare functions [14, 15]. Another line of emerging research, pluralistic alignment [16], seeks to reflect diverse perspectives in AI systems through approaches such as mixture-based models [17], belief-conditioned models [18], and steerable models [19], with a particular focus on LLMs. However, these methods generally assume explicit knowledge or clear labels of evaluator groups, which limits their practical applicability since group identities are often implicit or unobservable in real-world. Motivated by this limitation, our research aims to achieve proportional alignment without requiring additional information about the evaluator profile.

Our approach builds upon recent works addressing diverse preference aggregation through an axiomatic approach from social choice theory [12, 20–26]. Specifically, we propose a novel preference learning algorithm that satisfies two foundational axioms, monotonicity (ensuring that improving an alternative's ranking cannot decrease its probability) and Pareto efficiency (ensuring that if an alternative is preferred by all, it is favored by the policy), as well as two new axioms we introduce: population-proportional alignment (PPA) and population-bounded manipulability (PBM). The first new axiom, PPA, requires the policy to be at least weakly proportional to evaluator population shares, addressing RLHF and NLHF's insufficient representation of the population distribution of preferences. The second axiom, PBM, bounds the incentive for manipulation as an affine function of the true population share, thereby guaranteeing robustness. Recent studies have highlighted that conventional preference learning methods are susceptible to strategic misreporting [27]. Unlike existing approaches that incorporate explicit mechanism design to ensure strict strategyproofness [27–31], our method inherently limits manipulative advantage by constraining policy selection based on estimated feasible population distributions. Further details on related work are provided in Appendix B.

1.1 Our contribution

The first key contribution of this work is demonstrating that the set of feasible population distributions of evaluators can be inferred directly from pairwise comparison data. Leveraging this insight, we develop a novel preference learning framework designed to align policies proportionally with the underlying population distribution. To establish a rigorous theoretical basis, we adopt an axiomatic approach, proving that our framework satisfies two fundamental axioms, monotonicity and Pareto efficiency, and two newly introduced axioms, PPA and PBM. In addition, we propose a novel softmax relaxation method to control the trade-off between proportional alignment and the selection of the Condorcet winner. Finally, the proposed framework is validated through empirical evaluations in both tabular and function approximation settings.

Organization of the paper. In Section 2, we formalize the setting of preference learning and probabilistic social choice, and establish connections between them. In Section 3, motivated by a simple negative example, we introduce two desirable axioms alongside two fundamental axioms. In Section 4, we propose a novel preference learning algorithm that satisfies these axioms and provide a theoretical analysis. Finally, Section 5 presents empirical evaluations that demonstrate the effectiveness and scalability of our method. For ease of reference, all mathematical notation used in the paper is summarized in Appendix A.

2 Preliminaries

2.1 Probabilistic social choice function and preference learning

We begin by reviewing key concepts from social choice theory and preference learning to establish a foundation for our subsequent analysis. Consider a set of M alternatives, denoted by $\mathcal{Y} :=$

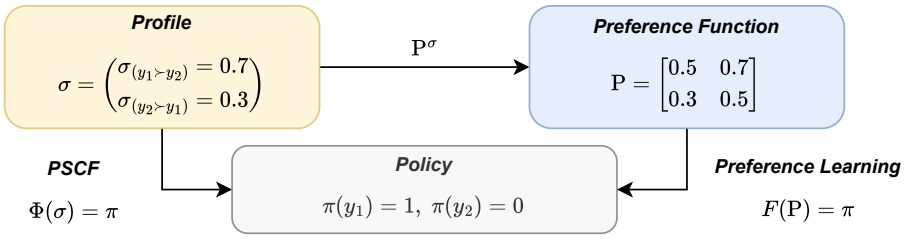

Figure 1: Illustration of the relationships between the profile, preference function, and policy.

$\{y_1, y_2, \ldots, y_M\}$, where each $y \in \mathcal{Y}$ may represent a response generated by a language model or an action in a decision-making task. We assume each evaluator has a strict and complete ranking over the alternatives, and let $\mathcal{S}$ denote the set of all possible rankings (i.e., permutations of $\mathcal{Y}$). Each ranking is represented by $r \in \mathcal{S}$, where $r(y_i) = k$ indicates that alternative y_i is ranked k-th under r. A *profile* $\sigma \in \Delta(\mathcal{S})$ is a probability distribution over the set of all rankings, where σ_r represents the proportion of the population that adheres to ranking r.

A *probabilistic social choice function* (PSCF) is a mapping $\Phi : \Delta(\mathcal{S}) \to \Delta(\mathcal{Y})$ that assigns to each profile σ a policy π, which is a probability distribution over the alternatives in $\mathcal{Y}$. In practice, however, acquiring a complete profile σ is often infeasible due to the high cost of collecting full rankings over a large set of alternatives.

To address this limitation, pairwise preference learning algorithms have been developed, allowing alignment based solely on pairwise comparison data. We define a *preference function* $\mathrm{P} : \mathcal{Y}^2 \to [0, 1]$, where $\mathrm{P}(y \succ y')$ denotes the probability that alternative y is preferred over y'. Given a profile σ, let P^σ be the preference function induced by the population distribution σ over rankings, defined as

$$\mathrm{P}^\sigma(y \succ y') := \sum_{r \in \mathcal{S}} \sigma_r \cdot \mathbf{1}_{\{r(y) < r(y')\}}, \tag{1}$$

where $\mathbf{1}_{r(y)<r(y')}$ is an indicator function equal to 1 if ranking r places alternative y in a better (i.e., lower) position than y', and 0 otherwise. This function captures the expected pairwise preference between y and y' under the distribution σ.

We define $\mathcal{P}$ as the set of all preference functions induced by some profile $\sigma \in \Delta(\mathcal{S})$:

$$\mathcal{P} := \{\mathrm{P} \mid \exists \sigma \in \Delta(\mathcal{S}) \text{ s.t. } \mathrm{P} = \mathrm{P}^\sigma\}. \tag{2}$$

Any $\mathrm{P} \in \mathcal{P}$ satisfies consistency conditions known as the Block-Marschak inequalities [32], including skew-symmetry: $\mathrm{P}(y \succ y') + \mathrm{P}(y' \succ y) = 1\ \forall y, y' \in \mathcal{Y}$. A *preference learning algorithm* is a mapping $F : \mathcal{P} \to \Delta(\mathcal{Y})$ that assigns a policy to each preference function. Throughout this paper, we say that a preference learning algorithm F *implements* a PSCF Φ if, for every profile $\sigma \in \Delta(\mathcal{S})$, it holds that $F(\mathrm{P}^\sigma) = \Phi(\sigma)$. The relationships between the profile, preference function, and policy are illustrated in Figure 1.

2.2 Two standard preference learning algorithms

Next, we introduce two prominent preference learning algorithms and discuss their connections to established concepts from probabilistic social choice theory.

Reinforcement learning from human feedback (RLHF). The Bradley–Terry (BT) model, widely used in preference modeling, assigns each alternative y_i a reward r_i with preference probabilities $\mathrm{P}(y_i \succ y_j) = \exp(r_i)/(\exp(r_i) + \exp(r_j))$. Standard RLHF estimates these rewards by likelihood maximization and then trains a policy to maximize expected rewards. Recent work [21] shows that this procedure is equivalent to the *maximal Borda rule* from social choice theory, which deterministically chooses the alternative with the highest Borda score $B(y) := \sum_{r \in \mathcal{S}} \sigma_r (M - r(y))$. As proved in Appendix C, the ranking from BT-optimized rewards coincides with Borda rankings, so RLHF without regularization (denoted by F^{RL}) implements the maximal Borda rule (denoted by Φ^{MB}). Direct preference optimization (DPO) [33] generalizes this by adding Kullback–Leibler (KL) regularization relative to a reference policy.

Nash learning from human feedback (NLHF). As highlighted in recent studies [9, 10, 12], RLHF has limitations in scenarios involving intransitive or cyclic preferences. An alternative y^* is called a *Condorcet winner* if it is preferred by a majority over every other alternative, formally stated as $\mathrm{P}(y^* \succ y) > 0.5$ for all $y \neq y^*$. When aggregating preferences across multiple evaluators, scenarios without a Condorcet winner can arise, which is called the *Condorcet paradox*. In such cases, selecting the alternative with the highest Borda score fails to adequately represent collective preferences, as a deterministic policy cannot capture the lack of consensus or nuanced preferences. To address intransitive preferences, the game-theoretic approach, known as Nash learning from human feedback (NLHF) [9–12], has been adopted to model preference learning as a two-player constant-sum game $\max_{\pi_1 \in \Delta(\mathcal{Y})} \min_{\pi_2 \in \Delta(\mathcal{Y})} \mathbb{E}_{(y_1,y_2)\sim(\pi_1,\pi_2)}[\mathrm{P}(y_1 \succ y_2)]$, where the equilibrium policy π^* cannot be uniformly outperformed. This algorithm, denoted by F^{NL}, implements the well-known PSCF *maximal lotteries (ML)* [34], denoted by Φ^{ML}.

3 Axiomatic framework for population-proportional alignment

3.1 Motivating example with binary alternatives

Despite their practical utility, neither RLHF nor NLHF guarantees alignment proportional to the evaluator's preferences. To illustrate this point, we present a simple scenario involving binary alternatives. Consider two alternatives, $\mathcal{Y} = \{y_1, y_2\}$, and a profile σ consisting of two distinct groups of evaluators: group G_1 prefers alternative y_1 over y_2, while group G_2 prefers y_2 over y_1. Let w_1^σ and w_2^σ denote the population shares of groups G_1 and G_2, respectively.

Suppose the two alternatives are nearly tied, with population distribution given by $(w_1^\sigma, w_2^\sigma) = (1/2 + \epsilon, 1/2 - \epsilon)$ for an arbitrarily small positive scalar ϵ. Then, the corresponding preference function is given by $\mathrm{P}^\sigma(y_1 \succ y_2) = 1/2+\epsilon$ and $\mathrm{P}^\sigma(y_2 \succ y_1) = 1/2-\epsilon$. Despite this minimal margin ϵ, both algorithms $F^{\mathrm{RL}}(\mathrm{P}^\sigma)$ and $F^{\mathrm{NL}}(\mathrm{P}^\sigma)$ yield a deterministic policy that select the alternative with slightly greater support, namely y_1.

This binary example highlights two potential limitations of RLHF and NLHF frameworks. First, selecting policies that focus entirely on a single alternative may not accurately represent preferences across evaluators, raising concerns about bias. Second, these methods have high sensitivity to small perturbations in preference function. Specifically, a slight shift in ϵ from negative to positive abruptly flips the policy outcome $(\pi(y_1), \pi(y_2))$ from $(0, 1)$ to $(1, 0)$, making such approaches vulnerable to small perturbations. These limitations underscore the need for a novel approach that reflects the ratio of (w_1^σ, w_2^σ) in the resulting policy.

3.2 Proposed Axioms for Population-Proportional Alignment and Robustness

Social choice theory studies the aggregation of individual preferences through an *axiomatic* approach, which specifies desirable properties (axioms) and characterizes aggregation rules that satisfy them. In particular, two fundamental axioms, *monotonicity* and *Pareto efficiency*, are presented in Appendix D. Following this approach, we introduce the axioms that a PSCF Φ is desired to satisfy and propose a preference learning algorithm F that implements such a PSCF.

Proposed axioms. Motivated by the earlier example, we next introduce a new axiom designed to ensure alignment with population distribution of preferences. Let $G_k := \{r \in \mathcal{S} \mid r(y_k) = 1\}$ denote the set of rankings in which alternative y_k is ranked first. The population share of group G_k is denoted by $w_k^\sigma := \sum_{r \in G_k} \sigma_r$. For notational convenience, we define $\sigma_k \in \Delta(\mathcal{S})$ as the normalized sub-profile restricted to rankings in G_k, where $\sigma_{k,r} = \sigma_r / w_k^\sigma$ for all $r \in G_k$, and $\sigma_{k,r} = 0$ for all $r \notin G_k$. Let P_k^σ denote the group-specific preference function, generated from σ_k, using the mapping defined in (1). By construction, $\mathrm{P}_k^\sigma(y_k \succ y) = 1$ for all $y \neq y_k$, since this group unanimously prefers y_k over all other alternatives. The overall preference function is then a weighted aggregation of the group-specific preferences: $\mathrm{P}^\sigma = \sum_{k=1}^M w_k^\sigma \mathrm{P}_k^\sigma$.

Under this definition, our first axiom ensures that the resulting policy proportionally reflects the underlying population shares of each group.

Table 1: Overview of standard PSCFs and axioms

Φ	F	Monotonicity	Pareto Efficiency	PPA	PBM
Maximal Borda (MB)	✓ (RLHF)	✓	✓	×	×
Maximal lotteries (ML)	✓ (NLHF)	×	✓	×	×
Random dictatorship (RD)	×	✓	✓	✓	✓
Proposed framework	✓	✓	✓	✓	✓

Definition 3.1 (α-Population-proportional alignment (α-PPA))**.** A PSCF Φ satisfies *α-population-proportional alignment* if $\pi(y_k)/w_k^\sigma \geq \alpha(\sigma)$ for all $\sigma \in \Delta(\mathcal{S})$ and $y_k \in \mathcal{Y}$, where $\pi = \Phi(\sigma)$ and $\alpha : \Delta(\mathcal{S}) \to (0, 1]$.

The function $\alpha(\sigma)$ quantifies the strength of alignment: a higher value of α implies stronger alignment with w^σ, with $\alpha(\sigma) = 1$ indicating perfect proportional alignment. Next, we examine the robustness of Φ against manipulation through the following axiom.

Definition 3.2 (Single-group manipulated profile)**.** Given a profile σ and a group index $k \in [M]$, a profile σ'_k is called a *single-group manipulated profile* of σ if σ'_k can be obtained by modifying only the ranking distribution of the sub-profile σ_k. Formally, σ'_k is a single-group manipulated profile of σ if there exists a profile σ' such that $\sigma'_k = \sigma + w_k^\sigma(\sigma' - \sigma_k)$.

Definition 3.3 (γ-Population-bounded manipulability (γ-PBM))**.** A PSCF Φ satisfies *γ-population-bounded manipulability* if, for any profile σ and its single-group manipulated profile σ'_k, we have $\Phi(\sigma'_k)(y_k) \leq \gamma_1 w_k^\sigma + \gamma_2$, where $\gamma = (\gamma_1, \gamma_2)$, $\gamma_1 > 0$, and $\gamma_1 + \gamma_2 = 1$.

The γ-PBM axiom ensures that the maximum influence a single group can exert through manipulation is bounded above by an affine function of its population share. Specifically, a group can only achieve a deterministic policy selection for its preferred alternative (i.e., $\Phi(\sigma'_k)(y_k) = 1$) only if it constitutes the entire evaluator population (i.e., $w_k^\sigma = 1$). Note that a larger γ_1 provides a stronger robustness guarantee. Particularly, $\gamma_1 = 1$ implies that the manipulated policy value is limited exactly to the group's true population share.

3.3 Limitations of Standard PSCFs: Axiom Violations and Non-Implementability

We next show that the standard PSCFs either fail to satisfy the proposed axioms or are not implementable by a preference learning algorithm. Consider a PSCF that aligns the policy exactly with each group's population distribution, commonly referred to as a *random dictatorship* [35].

Definition 3.4 (Random dictatorship)**.** A PSCF Φ^{RD} is called a *random dictatorship* if $\Phi^{\mathrm{RD}}(\sigma) = w^\sigma$ for all $\sigma \in \Delta(\mathcal{S})$.

By definition, Φ^{RD} satisfies both proposed axioms in their strongest forms: α-PPA with $\alpha(\sigma) = 1$ for all $\sigma \in \Delta(\mathcal{S})$, and γ-PBM with $\gamma = (1, 0)$. The following proposition establishes that Φ^{MB} and Φ^{ML} violate even the weakest forms of these axioms, whereas Φ^{RD} satisfies all four axioms.

Proposition 3.5. *Φ^{MB} and Φ^{ML} violate the α-PPA axiom for any α and the γ-PBM axiom for any γ. Φ^{RD} satisfies all four axioms.*

The proof is provided in Appendix E. Unfortunately, Φ^{RD} is not implementable by any pairwise preference learning algorithm, since distinct profiles σ_1 and σ_2 may induce identical preference functions $\mathrm{P}^{\sigma_1} = \mathrm{P}^{\sigma_2}$ but different population distributions $w^{\sigma_1} \neq w^{\sigma_2}$ (see Appendix F for an example). Because w^σ cannot be recovered solely from P^σ, no mapping from preference functions to policies can implement Φ^{RD}[3]. Our goal, therefore, is to construct a preference learning algorithm F that implements a PSCF Φ satisfying all four axioms. Table 1 summarizes the standard PSCFs, their implementability, and satisfaction of the four axioms; see Brandl et al. [37] for additional details.

[3] In the literature, the class of implementable PSCFs is often referred to as the C2 class [36]

4 Algorithmic framework and theoretical guarantees

4.1 Population distribution recovery from pairwise preferences

In this section, we introduce a preference algorithm F, which implements a PSCF satisfying all four axioms presented in the previous section. The framework first estimates the feasible set of underlying population distributions w from given pairwise preferences P, and subsequently constructs a policy π closely aligned with the inferred feasible set. We begin with the definition of a feasible population distribution and the characterization of the set of all feasible population distributions.

Definition 4.1. A population distribution w is considered *feasible* given P, if there exists a profile $\sigma \in \Delta(\mathcal{S})$ such that $w = w^\sigma$ and $\mathrm{P} = \mathrm{P}^\sigma$.

Proposition 4.2. *The set of all feasible population distributions given* P *can be expressed as*

$$\mathcal{W}(\mathrm{P}) := \Big\{ w \in \Delta(\mathcal{Y}) \;\Big|\; \exists (\mathrm{P}_1, \ldots, \mathrm{P}_M) \in \mathcal{P}^M \text{ s.t. } \mathrm{P} = \sum_{i=1}^{M} w_i \mathrm{P}_i, \\ \mathrm{P}_i(y_i \succ y) = 1 \; \forall y \in \mathcal{Y} \setminus \{y_i\},\; \forall i \in [M] \Big\}. \tag{3}$$

See Appendix G for the proof. In words, a population distribution w is feasible if and only if there exist group-specific preference functions $(\mathrm{P}_1, \ldots, \mathrm{P}_M)$ such that P is their weighted aggregation, and each P_i reflects a group of evaluators who unanimously prefer y_i over all other alternatives.

The exact characterization of the set $\mathcal{W}(\mathrm{P})$ is challenging due to the constraints imposed by the set $\mathcal{P}$. We therefore propose a tractable polyhedral outer approximation of the set $\mathcal{W}(\mathrm{P})$, with the number of constraints growing only linearly with the dimension M.

Definition 4.3. For each $i \in [M]$, define $u_i := \min_{y \in \mathcal{Y} \setminus \{y_i\}} \mathrm{P}(y_i \succ y)$.

Theorem 4.4. *The set of feasible population distributions satisfies*

$$\mathcal{W}(\mathrm{P}) \subseteq \overline{\mathcal{W}}(\mathrm{P}) := \Big\{ w \in \Delta(\mathcal{Y}) \;\Big|\; w_i \le u_i \; \forall i \in [M] \Big\}. \tag{4}$$

The proof is given in Appendix H. To provide intuition, note that $u_i = 1 - \max_{y \neq y_i} \mathrm{P}(y \succ y_i) = 1 - \mathrm{P}(y' \succ y_i)$, where y' is the alternative most preferred over y_i. Thus, u_i represents the remaining population share after excluding those who prefer y' to y_i. Thus, w_i cannot exceed this value, as the w_i proportion of evaluators would always report y_i as their preferred option. The tightness of the outer approximation is further discussed in Appendix I, and the relation between Theorem 4.4 and Tatli et al. [38] is examined in Appendix J.

Proposition 4.2 and Theorem 4.4 indicate a fundamental limitation in preference learning: since w^σ is not uniquely identifiable from pairwise comparison data, perfect PPA (i.e., $\alpha(\sigma) = 1$) is generally unattainable. This motivates designing algorithms that achieve α-PPA with the largest possible α.

4.2 Proposed algorithmic framework with axiomatic guarantees

Given a polyhedron $\overline{\mathcal{W}}(\mathrm{P})$, our goal is to select a policy π that guarantees the proportional alignment to all $w \in \overline{\mathcal{W}}(\mathrm{P})$. To this end, we propose to assign probabilities to alternatives in proportion to the derived upper bounds u_i.

Definition 4.5. The preference learning algorithm F^* maps a preference function P to the policy

$$\pi(y_i) = \frac{u_i}{\sum_{j=1}^{M} u_j} \quad \forall i \in [M]. \tag{5}$$

Let Φ^* denote the PSCF implemented by F^*.

This construction adopts a conservative strategy for handling uncertainty in w^σ by assigning probabilities proportional to the most conservative estimate of each w_i^σ. By doing so, the algorithm minimizes the worst-case misalignment caused by the inevitable information loss from pairwise comparisons. Formally, it solves $\max_{\pi \in \Delta(\mathcal{Y})} \min_{w \in \overline{\mathcal{W}}(\mathrm{P})} \|\pi / w\|_\infty$.

We first establish the foundational axiomatic guarantees of the proposed framework.

Theorem 4.6 (Monotonicity & Pareto efficiency)**.** *The proposed PSCF Φ^* satisfies the monotonicity and the Pareto efficiency.*

The proofs are provided in Appendix K. Next, we show that Φ^* satisfies the α-PPA axiom. The following lemma establishes that the ratio between the resulting policy and the true population share is lower bounded by the inverse of the total sum of the upper bounds u_i.

Lemma 4.7. *For any profile $\sigma \in \Delta(\mathcal{S})$, the policy $\pi = \Phi^*(\sigma)$ satisfies*

$$\frac{\pi(y_i)}{w_i^\sigma} \geq \left(\sum_{j=1}^{M} u_j \right)^{-1} \quad \forall i \in [M]. \tag{6}$$

The next lemma shows that this lower bound depends on the number of non-dominated alternatives:

Definition 4.8 (δ-dominated alternative)**.** For any $\delta \in [0,1]$, an alternative $y \in \mathcal{Y}$ is said to be *δ-dominated* in a profile σ if there exists an alternative $y' \in \mathcal{Y} \setminus \{y\}$ such that $\mathrm{P}^\sigma(y' \succ y) \geq \delta$.

Lemma 4.9. *Let $w^{\sigma,1}$ and $w^{\sigma,2}$ denote the largest and second-largest elements of w^σ, respectively. Consider any $\delta \in [0,1]$, and let N_δ^σ be the number of alternatives that are not δ-dominated in profile σ. Then the lower bound in* (6) *lies within the range $[\alpha(\sigma), 1]$, where*

$$\alpha(\sigma) := \left[(N_\delta^\sigma - 1)(1 - w^{\sigma,1}) + (1 - w^{\sigma,2}) + (M - N_\delta^\sigma)(1 - \delta) \right]^{-1}. \tag{7}$$

See Appendix L for the proofs. Combining both Lemmas, we obtain the following α-PPA guarantee:

Theorem 4.10 (α-PPA)**.** *The PSCF Φ^* satisfies the α-PPA axiom with α defined in* (7)*.*

Lemma 4.7 suggests that the actual alignment performance improves as $\sum_{j=1}^{M} u_j$ approaches 1. This typically occurs when the number of non-1-dominated alternatives is small. Notably, when there are only two non-1-dominated alternatives, substituting $N_1^\sigma = 2$ and $w^{\sigma,1} + w^{\sigma,2} = 1$ into (7) yields $\alpha(\sigma) = 1$, implying the perfect PPA in such cases. Moreover, when there exists a single dominating group, meaning $(w^{\sigma,1}, w^{\sigma,2})$ approaches $(1, 0)$, then $\alpha(\sigma)$ also approaches 1. Importantly, because $\sum_{j=1}^{M} u_j$ can be computed directly from a given preference function P, the alignment accuracy of the resulting policy can be evaluated at test time.

Finally, we present the population-bounded manipulability of the proposed method.

Theorem 4.11 (γ-PBM)**.** *Let $\pi' = \Phi^*(\sigma'_k)$ denote a policy resulting from single-group manipulation by group G_k. Then, the following inequality holds:*

$$\pi'(y_k) \leq \frac{u_k}{u_k + 1 - w_k^\sigma} \leq \frac{1}{2}(w_k^\sigma + 1). \tag{8}$$

Thus, the PSCF Φ^* satisfies γ-PBM with $(\gamma_1, \gamma_2) = (1/2, 1/2)$.

The proof is provided in Appendix M. Note that $(\gamma_1, \gamma_2) = (1/2, 1/2)$ represents a worst-case bound. The actual manipulability for each group is more tightly bounded by $u_k/(u_k + 1 - w_k^\sigma)$. For instance, if $u_k \leq 1/2$ and $w_k^\sigma \leq 1/2$, then $\pi'(y_k) \leq 1/2$. This indicates that a non-majority group cannot elevate their preferred alternative to majority status through manipulation. In addition, the above result can be extended to provide a weaker form of strategyproofness (see Appendix N for details).

4.3 Balancing PPA and Condorcet consistency

One potential limitation of the proposed framework is that Φ^* does not satisfy Condorcet consistency, a fundamental axiom which requires always selecting a Condorcet winner if one exists. In fact, it is impossible for any method to simultaneously satisfy both α-PPA and Condorcet consistency.

Definition 4.12 (Condorcet consistency)**.** A PSCF Φ satisfies *Condorcet consistency* if, for any profile σ with a Condorcet winner y^*, $\Phi(\sigma)(y^*) = 1$.

Proposition 4.13. *No PSCF can simultaneously satisfy α-PPA and Condorcet consistency.*

See Appendix O for the proof. To balance two axioms, we propose a softmax-relaxed algorithm F^β (and its corresponding PSCF Φ^β), by modifying F^* as follows:

$$\pi(y_i) = \frac{u_i \exp(\beta u_i)}{\sum_{j=1}^{M} u_j \exp(\beta u_j)} \quad \forall i \in [M]. \tag{9}$$

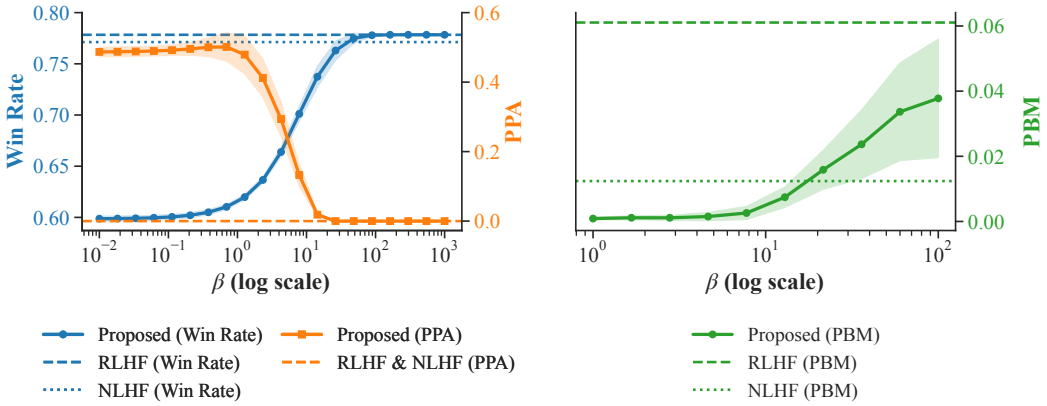

Figure 2: Tabular experiment results (Section 5.1) for F^{β}, F^{RL}, and F^{NL}. **Left:** win rate (left axis, blue) and PPA level (right axis, orange). **Right:** PBM level (policy gain through manipulation).

The parameter $\beta \geq 0$ controls how sharply the policy concentrates on alternatives with higher u_i values. When $\beta = 0$, the algorithm reduces to the original F^*. As $\beta \to \infty$, the policy becomes deterministic and converges to $\pi(y^*) = 1$, where $y^* = \arg\max_{i \in [M]} u_i$. This limiting Φ^{∞} is the well-known minimax Condorcet method [39], which satisfies Condorcet consistency (see Appendix P for the proof).

Proposition 4.14. Φ^{∞} *satisfies Condorcet consistency.*

The softmax relaxation offers a smooth trade-off between α-PPA and Condorcet consistency, controlled by the parameter β. We analyze the theoretical behavior of intermediate β values in Appendix Q, and empirically demonstrate the effects of varying β in Section 5. Additionally, Appendix R discusses the connection to pairwise majority consistency (PMC) [24], which imposes a stronger consistency requirement, ensuring the entire policy ranking aligns with majority preferences.

5 Experiments

5.1 Tabular experiment: movie recommendation

Datasets and experimental setup. To validate our theoretical findings, we evaluate the framework on a movie recommendation task using 1,297 evaluator rankings over 20 movies from MovieLens 1M dataset [40]. In each episode, we sample 10^5 pairwise comparisons i.i.d. from the true preference function P^{σ} and train F^{β} alongside two baselines, F^{RL} and F^{NL}.

We report averages and standard deviations over 50 episodes on three metrics: (i) win rate against a uniform policy, $\mathbb{E}_{(y_1,y_2)\sim(\pi,U)}[\mathrm{P}^{\sigma}(y_1 \succ y_2)]$, where U is the uniform distribution over $\mathcal{Y}$, (ii) PPA level, $\alpha(\sigma) = \min_{i \in [M]} \pi(y_i)/w_i^{\sigma}$, and (iii) PBM, the average policy gain from a single group's strategical manipulation.

Results and discussion. As shown in the left panel of Figure 2, RLHF and NLHF achieve high win rates of 0.7784 and 0.7712, respectively, but both yield a PPA level of 0. For our proposed algorithm F^{β}, we observe the expected trade-off: as β increases, the win rate rises from 0.5987 to 0.7784, while the PPA level decreases from 0.4869 to 0. These results confirm our theoretical prediction of each algorithm's behavior. Additionally, the average value of u_i was 0.1892, suggesting that the set $\overline{\mathcal{W}}(\mathrm{P})$ in (4) provides a meaningfully tight estimate of w^{σ} in our method.

Regarding PBM, the average gain was calculated as 0.0611 for RLHF, 0.0124 for NLHF, and 8.896×10^{-4} when $\beta = 10^0$. Overall, F^{β} outperforms the baselines when $\beta \leq 10^1$, indicating that our proposed algorithm significantly reduces susceptibility to manipulation and supports its robustness guarantee.

Table 2: Win rate and PPA level $\alpha(\sigma)$ across datasets and algorithms

Dataset	Category	Metric	$\beta = 0$	$\beta = 10^{-4}$	$\beta = 10^{-2}$	$\beta = 10^{0}$	DPO
Synthetic	Color	Win rate	0.6157	0.6880	0.6961	0.8429	**0.8566**
		PPA (α)	**0.0883**	0.0235	0.0183	0.0003	0.0000
Alpaca-GPT4	Expertise	Win rate	0.7613	0.7610	0.7634	0.7636	**0.7697**
		PPA (α)	**0.1428**	0.1418	0.1392	0.1273	0.1321
	Style	Win rate	0.8398	0.8432	0.8425	**0.8530**	0.8478
		PPA (α)	**0.5012**	0.4197	0.3637	0.3635	0.3786

5.2 Large-scale experiment: instruction-tuned LLMs

Datasets and experimental setup. We next evaluate the algorithm in high-dimensional settings with function approximation by fine-tuning the Qwen2.5-3B-Instruct model [41]. For a synthetic dataset, we construct 10 questions asking evaluators which color they prefer, with 10 candidate colors as possible responses. The true rankings of $1{,}000$ evaluators are generated from randomly sampled rewards, and 10^4 pairwise comparisons are drawn i.i.d. from P^{σ}. We next test the algorithm on the Alpaca-GPT4 dataset [42], which contains 52k prompts. Following prior work [8, 43], we consider two group categories (expertise and style) and sample one pairwise comparison per prompt using GPT-4.1 [44]. Further details on data generation and hyperparameters are provided in Appendix S.

For both datasets, we evaluate two metrics: (i) the win rate against a reference policy (the pretrained model), $\mathbb{E}_{(x,y_1,y_2)\sim(\rho,\pi,\pi_{\mathrm{ref}})}[\mathrm{P}^{\sigma}(y_1 \succ y_2 \mid x)]$, and (ii) the PPA level $\alpha(\sigma)$, comparing the results with DPO as the baseline. To estimate the output policy (i.e., the group distribution of generated responses), we used response logits directly for the synthetic dataset, and group classifications from the annotation model (GPT-4.1) for the Alpaca-GPT4 dataset. The specific training algorithm is described in Appendix T, and the full experimental code is included in the supplemental material.

Results and Discussion. Table 2 presents the win rate and PPA level $\alpha(\sigma)$ across datasets and algorithms. On the synthetic dataset, we observe a clear trade-off between win rate and PPA, confirming that β effectively controls this balance and validating the algorithm's effectiveness in high-dimensional settings. For the Alpaca-GPT4 dataset, a trade-off is observed, but the differences across β values are less distinct. This is because the group distributions of outputs are estimated using an annotation model (GPT-4.1), which introduces noise and obscures the effect of β. In contrast, the synthetic dataset allows direct computation from response logits, enabling more precise estimates.

We highlight several practical considerations for deployment. First, the current function approximation algorithm involves two-phase training (learning u and π), with computational costs comparable to RLHF and higher than DPO. This points to the need for direct policy optimization methods that balance win rate and proportional alignment. Second, accurately measuring the PPA level in LLMs remains an open challenge beyond the two approaches explored here (direct logit comparison and group classification). As this paper primarily introduces the theoretical framework with supporting experiments, our results should be viewed as encouraging evidence of scalability, and we expect that further work on refined algorithms and evaluation methods will strengthen these findings.

6 Conclusion and future directions

This paper introduces a novel preference-learning framework that aligns policies proportionally with population distributions inferred from pairwise comparison data. We believe this framework offers a new perspective on alignment algorithms by shifting the focus beyond the conventional emphasis on win rate. Furthermore, our work strengthens the connection between preference learning and social choice theory by implementing a new class of probabilistic social choice functions, extending beyond standard rules such as maximal Borda and maximal lotteries. Future research will aim to extend the framework to incorporate lower-ranked preferences and to develop more efficient algorithms for high-dimensional environments.

7 Acknowledgement

The authors acknowledge the MIT Office of Research Computing and Data for providing high performance computing resources that have contributed to the research results reported within this paper.

References

[1] Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback. In *Robotics Research: Volume 1*, pages 161–176. Springer, 2017.

[2] Erdem Bıyık, Nicolas Huynh, Mykel J Kochenderfer, and Dorsa Sadigh. Active preference-based gaussian process regression for reward learning. *arXiv preprint arXiv:2005.02575*, 2020.

[3] Wanqi Xue, Qingpeng Cai, Zhenghai Xue, Shuo Sun, Shuchang Liu, Dong Zheng, Peng Jiang, Kun Gai, and Bo An. Prefrec: Recommender systems with human preferences for reinforcing long-term user engagement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2874–2884, 2023.

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

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

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

[7] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[8] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[9] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 18, 2023.

[10] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

[11] Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807, 2024.

[12] Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. Jackpot! alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

[13] 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. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

[14] Huiying Zhong, Zhun Deng, Weijie J Su, Zhiwei Steven Wu, and Linjun Zhang. Provable multi-party reinforcement learning with diverse human feedback. *arXiv preprint arXiv:2403.05006*, 2024.

[15] Cheol Woo Kim, Jai Moondra, Shresth Verma, Madeleine Pollack, Lingkai Kong, Milind Tambe, and Swati Gupta. Navigating the social welfare frontier: Portfolios for multi-objective reinforcement learning. *arXiv preprint arXiv:2502.09724*, 2025.

[16] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

[17] Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.

[18] Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No preference left behind: Group distributional preference optimization. *arXiv preprint arXiv:2412.20299*, 2024.

[19] Jadie Adams, Brian Hu, Emily Veenhuis, David Joy, Bharadwaj Ravichandran, Aaron Bray, Anthony Hoogs, and Arslan Basharat. Steerable pluralism: Pluralistic alignment via few-shot comparative regression. *arXiv preprint arXiv:2508.08509*, 2025.

[20] Abhilash Mishra. Ai alignment and social choice: Fundamental limitations and policy implications. *arXiv preprint arXiv:2310.16048*, 2023.

[21] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

[22] Jessica Dai and Eve Fleisig. Mapping social choice theory to rlhf. *arXiv preprint arXiv:2404.13038*, 2024.

[23] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

[24] Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for ai alignment from human feedback. *arXiv preprint arXiv:2405.14758*, 2024.

[25] Zhekun Shi, Kaizhao Liu, Qi Long, Weijie J Su, and Jiancong Xiao. Fundamental limits of game-theoretic llm alignment: Smith consistency and preference matching. *arXiv preprint arXiv:2505.20627*, 2025.

[26] Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J Su. Theoretical tensions in rlhf: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*, 2025.

[27] Thomas Kleine Buening, Jiarui Gan, Debmalya Mandal, and Marta Kwiatkowska. Strategyproof reinforcement learning from human feedback. *arXiv preprint arXiv:2503.09561*, 2025.

[28] Chanwoo Park, Mingyang Liu, Dingwen Kong, Kaiqing Zhang, and Asuman Ozdaglar. Rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

[29] Ermis Soumalias, Michael J Curry, and Sven Seuken. Truthful aggregation of llms with an application to online advertising. *arXiv preprint arXiv:2405.05905*, 2024.

[30] Haoran Sun, Yurong Chen, Siwei Wang, Wei Chen, and Xiaotie Deng. Mechanism design for llm fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*, 2024.

[31] Shugang Hao and Lingjie Duan. Online learning from strategic human feedback in llm fine-tuning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[32] Henry David Block and Jacob Marschak. Random orderings and stochastic theories of response. *Cowles Foundation Discussion Papers*, 1959.

[33] 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. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[34] Peter C Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

[35] Felix Brandt. Rolling the dice: Recent results in probabilistic social choice. *Trends in computational social choice*, pages 3–26, 2017.

[36] Peter C Fishburn. Condorcet social choice functions. *SIAM Journal on applied Mathematics*, 33(3): 469–489, 1977.

[37] Florian Brandl, Felix Brandt, and Christian Stricker. An analytical and experimental comparison of maximal lottery schemes. *Social Choice and Welfare*, 58(1):5–38, 2022.

[38] Gokcan Tatli, Yi Chen, and Ramya Korlakai Vinayak. Learning populations of preferences via pairwise comparison queries. In *International Conference on Artificial Intelligence and Statistics*, pages 1720–1728. PMLR, 2024.

[39] Gerald H Kramer. A dynamical model of political equilibrium. *Cowles Foundation Discussion Papers*, 629, 1975.

[40] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[41] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[42] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[43] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[44] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[45] Kihyun Kim, Jiawei Zhang, Asuman Ozdaglar, and Pablo A Parrilo. A unified linear programming framework for offline reward learning from human demonstrations and feedback. *arXiv preprint arXiv:2405.12421*, 2024.

[46] Allan Gibbard. Manipulation of voting schemes: a general result. *Econometrica: journal of the Econometric Society*, pages 587–601, 1973.

A Notation

We summarize the mathematical notation used in the paper.

Symbol	Description
Rankings, profiles, groups, and preferences	
$[M]$	The set of integers $\{1, 2, \ldots, M\}$.
$\mathcal{Y} = \{y_1, \ldots, y_M\}$	The set of M alternatives.
$\Delta(\mathcal{Y})$	Probability simplex over a finite set $\mathcal{Y}$.
$\mathcal{S}$	The set of all rankings (permutations) over $\mathcal{Y}$.
$r \in \mathcal{S}$	A ranking, where $r(y_i) = k$ means y_i is ranked k-th.
$\sigma \in \Delta(\mathcal{S})$	A profile, i.e., population distribution over rankings.
$\sigma_r \in [0, 1]$	Proportion of evaluators who adopt ranking r.
G_k	Group k, set of rankings where y_k is ranked first.
$w_k^\sigma \in [0, 1]$	Population share of evaluators whose top choice is y_k.
$\sigma_k \in \Delta(\mathcal{S})$	Sub-profile of group G_k (evaluators who rank y_k first).
$\pi \in \Delta(\mathcal{Y})$	A policy, i.e., probability distribution over alternatives.
$\mathrm{P} \in \mathcal{P}$	Preference function, $\mathrm{P}(y \succ y')$ is the probability y is preferred to y'.
$\mathrm{P}^\sigma \in \mathcal{P}$	Preference function induced by a profile σ.
$\mathrm{P}_k, \mathrm{P}_k^\sigma$	Group-specific preference function for G_k.
$\mathcal{P}$	Set of all preference functions induced by some profile in $\Delta(\mathcal{S})$.
Preference learning algorithms, PSCFs, and axioms	
$F : \mathcal{P} \to \Delta(\mathcal{Y})$	Preference learning algorithm, mapping a preference function to a policy.
$\Phi : \Delta(\mathcal{S}) \to \Delta(\mathcal{Y})$	Probabilistic social choice function (PSCF), mapping a profile to a policy.
F^{RL}, Φ^{MB}	RLHF algorithm and its PSCF (maximal Borda rule).
F^{NL}, Φ^{ML}	NLHF algorithm and its PSCF (maximal lotteries).
$B(y)$	Borda score: $B(y) := \sum_{r \in \mathcal{S}} \sigma_r \, (M - r(y))$.
$\alpha(\sigma) \in \mathbb{R}$	Strength of population-proportional alignment (PPA) guarantee.
$\gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$	Parameters characterizing population-bounded manipulability (PBM).
σ'_k	Single-group manipulated profile of σ (group k perturbs only its sub-profile).
$u_i \in [0, 1]$	$u_i := \min_{y \neq y_i} \mathrm{P}(y_i \succ y)$, upper bound on feasible population share for y_i.
$\mathcal{W}(\mathrm{P})$	Set of feasible population distributions consistent with preference function P.
$\overline{\mathcal{W}}(\mathrm{P})$	Polyhedral outer approximation of $\mathcal{W}(\mathrm{P})$ (via $w_i \leq u_i$ constraints).
$\delta \in [0, 1]$	Dominance threshold (used in δ-domination definition).
N_δ^σ	Number of alternatives not δ-dominated under profile σ.
w_1^σ, w_2^σ	Largest and second-largest elements of w^σ.
y^*	Condorcet winner satisfying $\mathrm{P}(y^* \succ y) > \frac{1}{2}$ for all $y \neq y^*$.
F^*, Φ^*	Proposed (baseline) algorithm/PSCF using u_i with $\pi(y_i) \propto u_i$.
F^β, Φ^β	Softmax-relaxed algorithm/PSCF with concentration parameter $\beta \geq 0$.
Offline learning algorithm with function approximation	
$x \in \mathcal{X}$	Context (e.g., prompt or state) and context space.
$\mathcal{D} = \{(x_i, y_i^w, y_i^\ell)\}_{i=1}^N$	Offline dataset of pairwise comparisons (y^w preferred to y^ℓ).
$\rho(x), \pi_d(y \mid x)$	Context (prompt) and query data distribution.
μ	Selector model used to form u.
$\mathcal{F}_\mu$, $\mathcal{F}_\pi$	Function classes for μ and π.
$\widehat{\mathrm{P}}, \widehat{\mu}, \widehat{u}$	Empirical estimate of P, μ, and u.
$\widehat{\pi}_\beta, \widehat{\pi}$	Softmax policy constructed from $\widehat{u}$ and final estimated policy

B Additional related work

In this section, we discuss recent work that aims to capture and represent the diversity of human preferences in preference learning.

Limitations of the BT model. Recent studies have highlighted limitations of the standard RLHF approach under BT model assumption, which fails to capture the multifaceted and sometimes conflicting nature of human preferences. For example, Kim et al. [45] demonstrated that the standard MLE algorithms under the BT model can become unstable, particularly in the presence of evaluators exhibiting greedy behavior. They proposed to address this limitation by estimating a set of feasible reward functions without relying on specific modeling assumptions. Additionally, Siththaranjan et al. [21] established a theoretical equivalence between RLHF and the Borda voting rule, showing that the optimized rankings from standard methods frequently violate majority preferences. To address this issue, they introduced a distributional approach incorporating hidden context variables to address diverse evaluator preferences. Furthermore, Ge et al. [24] analyzed reward optimization methods under parameterizations, revealing their inherent violation of fundamental axioms such as Pareto efficiency. They proposed a novel algorithm explicitly designed to satisfy these axioms.

Approaches from social choice theory. Parallel research efforts have explored unbiased aggregation of heterogeneous human preferences, grounded in social choice theory. Chakraborty et al. [8] formally proved the impossibility of equitably aligning single-reward models across diverse evaluator groups, and proposed learning reward mixtures using the EM algorithm followed by maximizing the minimum utility across subpopulations. Additionally, Zhong et al. [14] conducted a rigorous analysis of multi-group reward learning under various social welfare criteria, such as Nash, utilitarian, and Leximin functions, and provided theoretical alignment guarantees. Park et al. [28] proposed a probabilistic opinion pooling function that directly aggregates multiple probabilistic models into a single policy, as well as personalized algorithms that output individualized policies after estimating confidence sets. Shi et al. [25] analyze the theoretical limits of NLHF, showing that exact preference matching is generally impossible, highlighting intrinsic limitations of this paradigm. Concurrent work by [26] is closely related to our work. They investigate the tension between RLHF's empirical success and its incompatibility with social choice axioms (PMC and Condorcet consistency), showing that RLHF can satisfy them under a practical assumption about preference labeling. Moreover, they propose a new axiom, *group preference matching*, which requires the policy to reproduce group-level preference distributions in proportion to their population weights. However, they do not provide an algorithmic framework that satisfies this axiom.

Pluralistic alignment. Emerging research on pluralistic alignment seeks to reflect diverse perspectives in AI systems, with a particular focus on LLMs. Sorensen et al. [16] outlined three complementary frameworks for pluralistic alignment: Overton pluralism, which captures the range of reasonable responses; steerable pluralism, which allows models to adapt to particular attributes; and distributional pluralism, which aligns model outputs with population-level distributions. Chen et al. [17] introduced a framework that modeled heterogeneous human preferences from the ground up using the ideal point model and mixture modeling. Yao et al. [18] proposed group distributional preference optimization (GDPO), a method that aligns models with the group preferences by estimating the underlying belief distribution and conditioning responses on those beliefs, ensuring representation of both majority and minority views. Adams et al. [19] developed a steerable pluralistic alignment algorithm, enabling models to adapt to individual preference profiles through few-shot comparative regression across fine-grained attributes. While these approaches show promise, they generally rely on explicit group identification, restricting their applicability in scenarios where group labels are unavailable or difficult to determine. In contrast, our work does not require explicit knowledge of evaluator groups. Instead, we infer population distributions directly from pairwise comparison data and align policies accordingly.

C Equivalence of BT-MLE rewards ranking and Borda ranking

Proposition C.1. *Let $r^* \in \mathbb{R}^M$ be a maximizer of the likelihood function*

$$L(r) := \sum_{i<j} \left[\mathrm{P}^{\sigma}(y_i \succ y_j) \log \left(\frac{e^{r_i}}{e^{r_i} + e^{r_j}} \right) + \mathrm{P}^{\sigma}(y_j \succ y_i) \log \left(\frac{e^{r_j}}{e^{r_i} + e^{r_j}} \right) \right]. \tag{10}$$

Then, the ordering of alternatives induced by r^* is identical to the ordering induced by the Borda score B of σ. Formally, for any $i, j \in [M]$,

$$r_i^* > r_j^* \quad \iff \quad B(y_i) > B(y_j). \tag{11}$$

Proof. The gradient of $L(r)$ with respect to r_i is given by:

$$\frac{\partial L(r)}{\partial r_i} = \sum_{j \neq i} \left[\mathrm{P}^\sigma(y_i \succ y_j) - \mathrm{sigmoid}(r_i - r_j) \right], \tag{12}$$

where $\mathrm{sigmoid}(x) := 1/(1 + e^{-x})$. At the optimal solution r^*, the first-order condition requires that

$$\sum_{j \neq i} \left[\mathrm{P}^\sigma(y_i \succ y_j) - \mathrm{sigmoid}(r_i^* - r_j^*) \right] = 0. \tag{13}$$

Now, consider two distinct alternatives i and k, and suppose that $r_i^* > r_k^*$. Since the sigmoid function is monotonically increasing, for any $j \neq i, k$, we have $\mathrm{sigmoid}(r_i^* - r_j^*) > \mathrm{sigmoid}(r_k^* - r_j^*)$, and also $\mathrm{sigmoid}(r_i^* - r_k^*) > \mathrm{sigmoid}(r_k^* - r_i^*)$. From the first-order conditions at optimality, we have:

$$\sum_{j \neq i} \mathrm{P}^\sigma(y_i \succ y_j) = \sum_{j \neq i} \mathrm{sigmoid}(r_i^* - r_j^*) \ \text{ and } \ \sum_{j \neq k} \mathrm{P}^\sigma(y_k \succ y_j) = \sum_{j \neq k} \mathrm{sigmoid}(r_k^* - r_j^*). \tag{14}$$

Since $r_i^* > r_k^*$, it follows that

$$\sum_{j \neq i} \mathrm{sigmoid}(r_i^* - r_j^*) > \sum_{j \neq k} \mathrm{sigmoid}(r_k^* - r_j^*). \tag{15}$$

Therefore, we have

$$\sum_{j \neq i} \mathrm{P}^\sigma(y_i \succ y_j) > \sum_{j \neq k} \mathrm{P}^\sigma(y_k \succ y_j). \tag{16}$$

By definition, $\mathrm{P}^\sigma(y_i \succ y_j) = \sum_{r \in \mathcal{S}} \sigma_r \cdot \mathbf{1}_{\{r(y_i) < r(y_j)\}}$. Substituting this into the inequality above, we get

$$\sum_{r \in \mathcal{S}} \sigma_r \cdot \sum_{j \neq i} \mathbf{1}_{\{r(y_i) < r(y_j)\}} > \sum_{r \in \mathcal{S}} \sigma_r \cdot \sum_{j \neq k} \mathbf{1}_{\{r(y_k) < r(y_j)\}}. \tag{17}$$

Recall that the Borda score is defined as $B(y) := \sum_{r \in \mathcal{S}} \sigma_r \cdot (M - r(y))$, we can rewrite the inner sums in the inequality as:

$$\sum_{j \neq i} \mathbf{1}_{\{r(y_i) < r(y_j)\}} = (M - 1) - (r(y_i) - 1) = M - r(y_i), \tag{18}$$

and similarly,

$$\sum_{j \neq k} \mathbf{1}_{\{r(y_k) < r(y_j)\}} = M - r(y_k). \tag{19}$$

Thus, the inequality becomes

$$\sum_{r \in \mathcal{S}} \sigma_r \cdot (M - r(y_i)) > \sum_{r \in \mathcal{S}} \sigma_r \cdot (M - r(y_k)), \tag{20}$$

which is equivalent to $B(y_i) > B(y_k)$.

For the converse, assume $B(y_i) > B(y_k)$. Following similar steps in reverse, this implies

$$\sum_{j \neq i} \mathrm{P}^\sigma(y_i \succ y_j) > \sum_{j \neq k} \mathrm{P}^\sigma(y_k \succ y_j), \tag{21}$$

which leads to

$$\sum_{j \neq i} \mathrm{sigmoid}(r_i^* - r_j^*) > \sum_{j \neq k} \mathrm{sigmoid}(r_k^* - r_j^*). \tag{22}$$

This inequality can only hold if $r_i^* > r_k^*$. Therefore, we have shown that $r_i^* > r_j^* \iff B(y_i) > B(y_j)$, completing the proof. □

D Fundamental axioms: monotonicity and Pareto efficiency

In this section, we present the definition of two fundamental axioms in social choice theory: *monotonicity* and *Pareto efficiency*. For detailed discussions of these axioms, we refer readers to [24, 35].

Definition D.1 (Monotonicity)**.** A PSCF Φ satisfies monotonicity if, for any alternative $y \in \mathcal{Y}$, improving its ranking in a profile without changing other relative rankings cannot decrease its probability in the resulting policy. Formally, if profile σ' is obtained from σ by improving the ranking of y in some $r \in \mathcal{S}$ with $\sigma_r > 0$, then $\Phi(\sigma')(y) \geq \Phi(\sigma)(y)$.

Definition D.2 (Pareto efficiency)**.** A PSCF Φ satisfies *Pareto efficiency* if, whenever an alternative y is ranked above y' in every ranking r with nonzero population share, the resulting policy assigns at least as much probability to y as to y'. Formally, if $r(y) < r(y')$ for all $r \in \mathcal{S}$ with $\sigma_r > 0$, then $\Phi(\sigma)(y) \geq \Phi(\sigma)(y')$.

E Proof of Proposition 3.5

We first demonstrate that Φ^{MB} and Φ^{ML} violate α-PPA. Consider a preference profile σ with the following characteristics: (i) The population share of each group is nearly identical, with G_1 having a population share w_1^σ that is ϵ greater than the average, and G_2 having a population share w_2^σ that is ϵ less than the average. (ii) Within each group G_k, there is indifference between any two alternatives other than y_k. That is, $\mathrm{P}_k^\sigma(y_i \succ y_j) = 1/2$ for all $i, j \neq k$. Given this profile, we will show that for any $\epsilon > 0$, both RLHF and NLHF yield a deterministic policy that selects y_1.

The population distribution and pairwise preference function satisfy

$$w^\sigma = \left(\frac{1}{M} + \epsilon, \frac{1}{M} - \epsilon, \frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}\right) \text{ and } \mathrm{P}_k^\sigma(y_i \succ y_j) = \frac{1}{2}, \quad \forall i, j \neq k. \tag{23}$$

Then, the aggregated pairwise preferences P^σ are computed as follows:

- $\mathrm{P}^\sigma(y_1 \succ y_2) = \frac{1}{2} + \epsilon$
- $\mathrm{P}^\sigma(y_1 \succ y) = \frac{1}{2} + \frac{\epsilon}{2}$ for any $y \neq y_1, y_2$
- $\mathrm{P}^\sigma(y_2 \succ y) = \frac{1}{2} - \frac{\epsilon}{2}$ for any $y \neq y_1, y_2$
- $\mathrm{P}^\sigma(y \succ y') = \frac{1}{2}$ for any $y, y' \neq y_1, y_2$

Under this profile, for any $\epsilon > 0$, both Φ^{MB} and Φ^{ML} result in a policy where $\pi(y_1) = 1$, and $\pi(y_i) = 0$ for any $i \neq 1$. This implies that $\pi(y_i)/w_i^\sigma = 0$ for any $i \neq 1$, which violates α-PPA for any $\alpha > 0$.

Next, we show that Φ^{ML} violates γ-PBM using the profile described earlier with $M = 3$. The aggregated preference function P^σ can be represented by the following matrix:

$$\mathrm{P}^\sigma = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} + \epsilon & \frac{1+\epsilon}{2} \\ \frac{1}{2} - \epsilon & \frac{1}{2} & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \frac{1+\epsilon}{2} & \frac{1}{2} \end{bmatrix}. \tag{24}$$

Now, suppose that group G_3 manipulates their sub-profile from $\mathrm{P}_3^\sigma(y_1 \succ y_2) = \frac{1}{2}$ to $\mathrm{P}_3^{\sigma'}(y_1 \succ y_2) = 0$. Then, the resulting manipulated aggregated preference function $\mathrm{P}^{\sigma'}$ is calculated as:

$$\mathrm{P}^{\sigma'} = \begin{bmatrix} \frac{1}{2} & \frac{1}{3} + \epsilon & \frac{1+\epsilon}{2} \\ \frac{2}{3} - \epsilon & \frac{1}{2} & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \frac{1+\epsilon}{2} & \frac{1}{2} \end{bmatrix}. \tag{25}$$

Φ^{ML} yields a stochastic policy that depends on the value of ϵ. For example, if $\epsilon = 1/12$, the resulting policy is $\pi = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{2}\right]$. However, as ϵ approaches 0, the resulting policy converges to $[0, 0, 1]$. This shows that $\pi'(y_3) \to 1$ while $w_3^\sigma = 1/3$, thus demonstrating that there exists no $\gamma_1 > 0$ for which Φ^{ML} satisfies γ-PBM.

To show that Φ^{MB} violates γ-PBM, consider the case with $M = 3$ where the profile σ consists of the following three groups of evaluators:

$$\sigma = \{(y_1 \succ y_2 \succ y_3) \times 0.30,\ (y_2 \succ y_1 \succ y_3) \times 0.45,\ (y_3 \succ y_1 \succ y_2) \times 0.25\}, \tag{26}$$

where $(y_1 \succ y_2 \succ y_3)$ represents a ranking r and "$\times 0.30$" indicates that $\sigma_r = 0.30$. Then, the Borda scores are calculated as $B = [1.3, 1.20, 0.5]$. Thus, $\Phi^{\mathrm{MB}}(\sigma) = \pi$, where $\pi(y_1) = 1$. Next, suppose the second group strategically misreports their preference from $(y_2 \succ y_1 \succ y_3)$ to $(y_2 \succ y_3 \succ y_1)$. Then, the Borda scores are calculated as $B' = [0.85, 1.2, 0.95]$. The resulting policy is then $\pi'(y_2) = 1$, with the population share of the second group being $w_2^\sigma = 0.45$. This example demonstrates that there exists no $\gamma_1 > 0$ for which Φ^{MB} satisfies γ-PBM.

Next, we show that Φ^{RD} satisfies all four axioms. Φ^{RD} satisfies monotonicity because improving ranking of y cannot decrease the number of evaluators whose top choice is y. In addition, Φ^{RD} satisfies Pareto efficiency because if $r(y_j) < r(y_k)$ for all $r \in \mathcal{S}$ with $\sigma_r > 0$, then we have $w_k^\sigma = 0$ and $\Phi^{\mathrm{RD}}(\sigma)(y_k) = 0$. Additionally, Φ^{RD} satisfies α-PPA with $\alpha(\sigma) = 1$ for all σ by its definition, and also satisfy γ-PBM with $(\gamma_1, \gamma_2) = (1, 0)$ because each group G_k cannot increase w_k^σ by manipulation.

F Proof of the non-implementability of Φ^{RD}

Suppose that Φ^{RD} can be implemented by a preference learning algorithm F^{RD}. Let $M = 3$, and consider two preference profiles, σ_1 and σ_2, defined as follows:

$$\begin{aligned} \sigma_1 &= \{(y_1 \succ y_2 \succ y_3) \times 1/3,\ (y_2 \succ y_1 \succ y_3) \times 1/3,\ (y_3 \succ y_1 \succ y_2) \times 1/3\}, \\ \sigma_2 &= \{(y_1 \succ y_2 \succ y_3) \times 2/3,\ (y_3 \succ y_2 \succ y_1) \times 1/3\}. \end{aligned} \tag{27}$$

Both of these profiles induce the same aggregated preference function $\mathrm{P}^\sigma = \mathrm{P}^{\sigma_1} = \mathrm{P}^{\sigma_2}$, where

$$\mathrm{P}^\sigma = \begin{bmatrix} \frac{1}{2} & \frac{2}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{2} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{2} \end{bmatrix}. \tag{28}$$

Therefore, the preference learning algorithm F^{RD} would produce the same policy for both σ_1 and σ_2. However, according to the definition of Φ^{RD}, we have $\Phi^{\mathrm{RD}}(\sigma_1) = [1/3, 1/3, 1/3]$ and $\Phi^{\mathrm{RD}}(\sigma_2) = [2/3, 0, 1/3]$, which are different policies. This implies that F^{RD} does not implement Φ^{RD}, which contradicts our initial assumption. Therefore, Φ^{RD} is not implementable by a preference learning algorithm.

G Proof of Proposition 4.2

First, consider any feasible population share w given a preference function P. By Definition 4.1, there exists a profile σ such that $w = w^\sigma$ and $\mathrm{P} = \mathrm{P}^\sigma$. Then, the group-specific preference functions $(\mathrm{P}_1^\sigma, \ldots, \mathrm{P}_M^\sigma)$ that constitute P^σ, satisfy the condition in (3), which implies that $w \in \mathcal{W}(\mathrm{P})$.

Next, consider any $w \in \mathcal{W}(\mathrm{P})$. By the definition of $\mathcal{W}(\mathrm{P})$, there exist $(\mathrm{P}_1, \ldots, \mathrm{P}_M) \in \mathcal{P}^M$ such that $\mathrm{P} = \sum_{k=1}^M w_k \mathrm{P}_k$, where each P_k satisfies $\mathrm{P}_k(y_k \succ y) = 1$ for all $y \neq y_k$. Since $\mathrm{P}_k \in \mathcal{P}$, there exists a profile σ_k that induces P_k, such that $\mathrm{P}_k = \mathrm{P}^{\sigma_k}$. Now, if we consider an aggregated profile $\sigma := \sum_{k=1}^M w_k \sigma_k$ by combining these group profiles with the corresponding weights, then the preference function of σ will be $\mathrm{P}^\sigma = \sum_{k=1}^M w_k \mathrm{P}^{\sigma_k} = \sum_{k=1}^M w_k \mathrm{P}_k = \mathrm{P}$ and also $w^\sigma = w$. Therefore, w is a feasible population distribution given P.

H Proof of Theorem 4.4

Consider any $w \in \mathcal{W}(\mathrm{P})$. Then, there exists $(\mathrm{P}_1, \ldots, \mathrm{P}_M) \in \mathcal{P}^M$ such that $\mathrm{P} = \sum_{k=1}^M w_k \, \mathrm{P}_k$. Fix an index $i \in [M]$. For any $y \in \mathcal{Y} \setminus \{y_i\}$, we have

$$\mathrm{P}(y_i \succ y) = w_i \, \mathrm{P}_i(y_i \succ y) + \sum_{k \neq i} w_k \, \mathrm{P}_k(y_i \succ y) \geq w_i \tag{29}$$

since $\mathrm{P}_i(y_i \succ y) = 1$. Taking the minimum over $y \in \mathcal{Y} \setminus \{y_i\}$ yields

$$\min_{y \in \mathcal{Y} \setminus \{y_i\}} \mathrm{P}(y_i \succ y) \geq w_i, \tag{30}$$

which implies $w_i \leq u_i \ \forall i \in [M]$ and $w \in \overline{\mathcal{W}}(\mathrm{P})$. Therefore, $\mathcal{W}(\mathrm{P}) \subseteq \overline{\mathcal{W}}(\mathrm{P})$.

I Additional Remarks on the Tightness of the Outer Approximation

The gap between the true feasible set $\mathcal{W}(\mathrm{P})$ and its outer approximation $\overline{\mathcal{W}}(\mathrm{P})$ arises from our profile assumption, namely that each evaluator has a strict and complete ranking. To illustrate this point, we show that $\overline{\mathcal{W}}(\mathrm{P})$ provides a tight approximation (i.e., $\mathcal{W}(\mathrm{P}) = \overline{\mathcal{W}}(\mathrm{P})$) under an extended profile setting. Consider an extended profile setting in which each group G_k is allowed to provide pairwise comparison data according to its own preference function P_k, subject only to the skew-symmetry constraint $\mathrm{P}_k(y_i \succ y_j) + \mathrm{P}_k(y_j \succ y_i) = 1$ for all $y_i, y_j \in \mathcal{Y}$, and the unanimity constraint $\mathrm{P}_k(y_k \succ y) = 1$ for all $y \in \mathcal{Y}$. In this case, the set $\mathcal{P}$ is defined as

$$\mathcal{P} := \{\mathrm{P} \mid \mathrm{P}(y \succ y') + \mathrm{P}(y' \succ y) = 1 \ \forall y, y' \in \mathcal{Y}\}. \tag{31}$$

We show that $\overline{\mathcal{W}}(\mathrm{P}) \subseteq \mathcal{W}(\mathrm{P})$ also holds under this setting. Consider any $w \in \overline{\mathcal{W}}(\mathrm{P})$. By assumption, w satisfies $w_i \leq u_i = \min_{y \in \mathcal{Y} \setminus \{y_i\}} \mathrm{P}(y_i \succ y)$ for all $i \in [M]$. Define each element of $(\mathrm{P}_1, \ldots, \mathrm{P}_M)$ as

$$\mathrm{P}_k(y_i \succ y_j) = \frac{\mathrm{P}(y_i \succ y_j) - w_i}{1 - w_i - w_j} \tag{32}$$

for any $i, j \neq k$, and let $\mathrm{P}_k(y_k \succ y) = 1$, $\mathrm{P}_k(y \succ y_k) = 0$ for all $y \in \mathcal{Y}$. Then $\mathrm{P}_k(y_i \succ y_j) \in [0, 1]$ holds because $\mathrm{P}(y_i \succ y_j) \in [w_i, 1 - w_j]$ by assumption. The skew-symmetry condition $\mathrm{P}_k(y_i \succ y_j) + \mathrm{P}_k(y_j \succ y_i) = 1$ is also satisfied. Thus, $\mathrm{P}_k \in \mathcal{P}$ and P_k can be induced by some profile. Finally, the constraint $\mathrm{P} = \sum_{k=1}^{M} w_k \mathrm{P}_k$ also holds. Therefore, $w \in \mathcal{W}(\mathrm{P})$, implying $\overline{\mathcal{W}}(\mathrm{P}) \subseteq \mathcal{W}(\mathrm{P})$, and hence $\mathcal{W}(\mathrm{P}) = \overline{\mathcal{W}}(\mathrm{P})$.

J Connection of Theorem 4.4 and Tatli et al. [38]

Tatli et al. [38] studies the recovery of population preference distributions under a spatial model. In their framework, each alternative is represented by a feature vector in a Euclidean space, and each voter's preferences are determined by distances to these vectors (i.e., voters prefer alternatives that are closer in the Euclidean norm). Theorem 4.4 can also be derived in this setting. Specifically, consider a sufficiently high-dimensional feature space partitioned into $M!$ regions by $\binom{M}{2}$ hyperplanes, where each hyperplane is the perpendicular bisector of the line segment connecting a pair of alternative vectors. Then, [38, Proposition 2] shows that it is impossible to recover the full profile σ from aggregated pairwise comparison data P^σ. Moreover, by summing the inequality in [38, Proposition 4] over the regions corresponding to voters who most prefer each alternative y_i, we obtain the bound $w_i \leq u_i$.

K Proof of Theorem 4.6

We first prove monotonicity. Improving the ranking of y_i for some evaluator can only increase $\mathrm{P}^\sigma(y_i \succ y)$ for any $y \neq y_i$, and decrease $\mathrm{P}^\sigma(y \succ y_i)$. This implies that u_i cannot decrease, while u_j for $j \neq i$ cannot increase. Therefore, $\pi(y_i) = u_i / (\sum_{j=1}^{M} u_j)$ cannot decrease, establishing monotonicity.

Next, we prove Pareto efficiency. Suppose y_i is ranked above y_j in every input ranking, i.e., $r(y_i) < r(y_j)$ for all $r \in \mathcal{S}$ with $\sigma_r > 0$. Then, we have $\mathrm{P}^\sigma(y_i \succ y_j) = \sum_{r \in \mathcal{S}} \sigma_r \cdot \mathbf{1}_{\{r(y_i) < r(y_j)\}} = 1$ and also $\mathrm{P}^\sigma(y_j \succ y_i) = 0$. Thus, we get $u_j = \min_{y \in Y \setminus \{y_j\}} \mathrm{P}^\sigma(y_j \succ y) = 0$. Thus, the resulting policy satisfies $\Phi^*(\sigma)(y_j) = u_j / (\sum_{k=1}^{M} u_k) = 0$. Therefore, $\Phi^*(\sigma)(y_i) \geq \Phi^*(\sigma)(y_j)$, establishing that Φ^* satisfies Pareto efficiency.

L Proof of Lemma 4.7 and Lemma 4.9

Lemma 4.7 follows directly from the fact that $w_i^\sigma \leq u_i$ for all $i \in [M]$, which gives

$$\frac{\pi(y_i)}{w_i^\sigma} = \frac{u_i}{w_i^\sigma \sum_{j=1}^{M} u_j} \geq \frac{1}{\sum_{j=1}^{M} u_j}. \tag{33}$$

Next, we show Lemma 4.9. Let $I \subseteq [M]$ be the set of indexes for δ-dominated alternatives, where $|I| = N_\delta^\sigma$. Then, for any $i \in I$, we have

$$u_i = \min_{j \in [M] \setminus \{i\}} \mathrm{P}^\sigma(y_i \succ y_j) \leq \mathrm{P}^\sigma(y_i \succ y') \leq 1 - \delta, \tag{34}$$

where y_i' denotes an alternative that δ-dominates y_i. Additionally, let $k \in \arg\max_{i \in [M]} w_i^\sigma$. Then, for any $i \neq k$, we have

$$u_i = \min_{j \in [M] \setminus \{i\}} \mathrm{P}^\sigma(y_i \succ y_j) \leq \mathrm{P}^\sigma(y_i \succ y_k) \leq 1 - w_k^\sigma \tag{35}$$

Similarly, let $l \in \arg\max_{i \in [M], i \neq k} w_i^\sigma$, then we have $u_k \leq 1 - w_l^\sigma$. Combining these reesults,

$$\sum_{i=1}^{M} u_i = u_k + \sum_{i \neq k, i \notin I} u_i + \sum_{i \in I} u_i \leq (1 - w_l^\sigma) + (N_\delta^\sigma - 1)(1 - w_k^\sigma) + (M - N_\delta^\sigma)(1 - \delta). \tag{36}$$

In addition, since $u_i \geq w_i^\sigma$, we have $\sum_{i=1}^M u_i \geq \sum_{i=1}^M w_i^\sigma = 1$. Combining both inequalities and plugging $(w_k^\sigma, w_l^\sigma) = (w^{\sigma,1}, w^{\sigma,2})$ in, we get the result of Lemmma 4.9 as follows:

$$\left(\sum_{i=1}^{M} u_i\right)^{-1} \in \left[\frac{1}{(N_\delta^\sigma - 1)(1 - w^{\sigma,1}) + (1 - w^{\sigma,2}) + (M - N_\delta^\sigma)(1 - \delta)},\ 1\right]. \tag{37}$$

M Proof of Theorem 4.11

Let σ' be the profile manipulated by group G_k, and let $\pi' = \Phi^*(\sigma')$ be the resulting policy. G_k aims to maximize

$$\pi'(y_k) = \frac{u_k'}{u_k' + \sum_{i \neq k} u_i'}, \tag{38}$$

where u' represents the value of u after the manipulation. To maximize $\pi'(y_k)$, G_k will attempt to maximize u_k' and minimize $\sum_{i \neq k} u_i'$. Since increasing the ranking of y_k in their profile increases (or at least does not decrease) the value of u_k' without increasing the value of $\sum_{i \neq k} u_i'$, the optimal strategy for G_k is to truthfully report y_k as its top choice. In this strategy, we have $u_k' = u_k$ and the sum $\sum_{i \neq k} u_i'$ has the following lower bound:

$$\sum_{i \neq k} u_i' \geq \sum_{i \neq k} w_i^\sigma = 1 - w_k^\sigma. \tag{39}$$

Substituting this lower bound into (38), we obtain

$$\pi'(y_k) \leq \frac{u_k}{u_k + 1 - w_k^\sigma} \leq \frac{1}{2}(w_k^\sigma + 1), \tag{40}$$

where the final inequality holds if $(w_k^\sigma - u_k + 1)(w_k^\sigma - 1) \leq 0$, which follows from the fact that $w_k^\sigma, u_k \in [0, 1]$.

N Population-bounded strategyproofness guarantee

In social choice theory, a mechanism is considered strategyproof if participants cannot benefit (i.e., increase their utility) by misreporting their true preferences [46], regardless of what other participants report. In our preference learning framework, we assume each group G_k's utility is the probability assigned to its top choice, represented by $\pi(y_k)$. A preference learning algorithm is strategyproof if no group can improve its outcome by misreporting preferences. However, as noted by Buening et al. [27], strict strategyproofness is typically too restrictive and is not satisfied by the conventional preference learning algorithms. Our method does not satisfy strict strategyproofness like other methods, but satisfies a weaker form that provides a bounded guarantee on the maximum potential gain from strategic misreporting in equilibrium.

Let σ' denote the profile resulting from strategical misreporting by all groups, and let $\pi' = \Phi^*(\sigma')$ be the resulting policy. Each group G_k aims to maximize $\pi'(y_k)$, which involves maximizing u_k' and minimizing $\sum_{i \neq k} u_i'$.

Since improving the ranking of y_k in their reported preferences increases (or at worst, does not decrease) the value of u'_k without increasing $\sum_{i \neq k} u'_i$, the optimal strategy for G_k is to truthfully report y_k as their top choice. Hence, all groups truthfully report their top choice regardless of other groups' strategies, meaning $\mathrm{P}'_k(y_k \succ y) = 1$ for all $y \neq y_k$, where P'_k denotes the reported preference function of G_k.

In this equilibrium, following steps analogous to the proof of Theorem 4.11, we have:

$$\pi'(y_k) \leq \frac{u'_k}{u'_k + 1 - w^\sigma_k} \leq \frac{1}{2}(w^\sigma_k + 1) \quad \forall k \in [M]. \tag{41}$$

Note that $\gamma(w^\sigma_k)$ is not a tight bound. Further exploration into tighter bounds and detailed analysis of each group's strategic behavior is left for future research.

O Proof of Proposition 4.13

Suppose $M = 2$ and $\mathrm{P}^\sigma(y_1 \succ y_2) \in (0.5, 1)$, so y_1 is the Condorcet winner. If a PSCF Φ satisfies Condorcet consistency, it must return the deterministic policy $\pi(y_1) = 1$. However, this violates the α-PPA axiom because $\pi(y_2) = 0$ while $w^\sigma_2 > 0$, which implies that $\pi(y_2)/w^\sigma_2 = 0$ cannot be lower bounded by any $\alpha(\sigma) > 0$.

P Proof of Proposition 4.14

Suppose y_i is a Condorcet winner. Then $\mathrm{P}^\sigma(y_i \succ y_j) > 0.5$ for all $j \neq i$, which implies that $u_i > 0.5$. For any other $j \neq i$, we have $u_j \leq \mathrm{P}^\sigma(y_j \succ y_i) < 0.5$. Therefore, y_i has the highest u_i, i.e., $i \in \arg\max_{j \in [M]} u_j$, and Φ^∞ returns $\pi(y_i) = 1$, satisfying Condorcet consistency.

Q Finite Behavior of Φ^β

The following proposition quantifies how large the parameter β needs to be to ensure that a Condorcet winner receives a sufficiently high probability under the softmax policy.

Proposition Q.1 (Condorcet consistency at finite β)**.** *Let y_i be a Condorcet winner with $u_i > 0.5$. Then, the softmax policy satisfies $\pi(y_i) \geq \alpha_c$ if*

$$\beta \geq \frac{1}{u_i - 0.5} \log\left(\frac{(M-1)\alpha_c}{2(1-\alpha_c)}\right). \tag{42}$$

Proof. Since y_i is a Condorcet winner, we have $u_j \leq \mathrm{P}^\sigma(y_j \succ y_i) = 1 - \mathrm{P}^\sigma(y_i \succ y_j) < 0.5$ for any $j \neq i$. From the given condition

$$\beta \geq \frac{1}{u_i - 0.5} \log\left(\frac{(M-1)\alpha_c}{2(1-\alpha_c)}\right), \tag{43}$$

we can establish the following lower bound:

$$u_i \exp(\beta u_i) \geq \frac{\alpha_c}{1-\alpha_c}(M-1)(0.5\exp(0.5\beta)) \geq \frac{\alpha_c}{1-\alpha_c}\sum_{j \neq i} u_j \exp(\beta u_j). \tag{44}$$

Thus, the softamx policy satisfies

$$\pi(y_i) = \frac{u_i \exp(\beta u_i)}{u_i \exp(\beta u_i) + \sum_{j \neq i} u_j \exp(\beta u_j)} \geq \alpha_c. \tag{45}$$

□

In addition, it can be shown that the α-PPA guarantee deteriorates as $\beta \to \infty$, since the lower bound in Lemma 4.7 becomes

$$\frac{\pi(y_i)}{w^\sigma_i} \geq \left(\sum_{j=1}^{M} u_j \exp(\beta(u_j - u_i))\right)^{-1}, \tag{46}$$

which converges to zero as $\beta \to \infty$, unless $u_i = \max_{j \in [M]} u_j$.

R Connection to pairwise majority consistency (PMC)

We discuss the connection to pairwise majority consistency (PMC) [24], which imposes a stronger consistency requirement, ensuring the entire policy ranking aligns with majority preferences.

Definition R.1 (Pairwise majority consistent ranking (PMC ranking))**.** A ranking r^σ is a called a *PMC ranking* of a profile σ if for all $y_i, y_j \in \mathcal{Y}$, a majority of evaluators prefer alternative y_i to alternative y_j in σ if and only if y_i is ranked higher than y_j in r^σ. Formally, $\mathrm{P}^\sigma(y_i \succ y_j) > 1/2$ if and only if $r^\sigma(y_i) < r^\sigma(y_j)$.

Definition R.2 (Pairwise majority consistency (PMC))**.** A PSCF Φ satisfies PMC if, for any profile $\sigma \in \Delta(\mathcal{S})$ that has a PMC ranking r^σ, $\Phi(\sigma)$ has the same ranking with r^σ, i.e. $\Phi(\sigma)(y_i) \geq \Phi(\sigma)(y_j)$ if $r^\sigma(y_i) < r^\sigma(y_j)$.

It can be shown that any Φ^β with finite $\beta \geq 0$ violates PMC, and only the limiting PSCF Φ^β satisfies PMC.

Proposition R.3. *Any Φ^β with finite $\beta \geq 0$ violates PMC. Φ^∞ satisfies PMC.*

Proof. First, we show that Φ^β violates PMC for any $\beta \geq 0$. It suffices to demonstrate that there exists a profile σ with a PMC ranking r^σ for which $\Phi^\beta(\sigma)(y_i) < \Phi^\beta(\sigma)(y_j)$ while $r^\sigma(y_i) < r^\sigma(y_j)$.

Consider the following profile σ with $M = 3$:

$$\sigma = \{(y_1 \succ y_2 \succ y_3) \times 0.3,\ (y_2 \succ y_3 \succ y_1) \times 0.1,\ (y_3 \succ y_1 \succ y_2) \times 0.3,\ (y_3 \succ y_2 \succ y_1) \times 0.3\}, \tag{47}$$

which yields the following preference function:

$$\mathrm{P}^\sigma = \begin{bmatrix} 0.5 & 0.6 & 0.3 \\ 0.4 & 0.5 & 0.4 \\ 0.7 & 0.6 & 0.5 \end{bmatrix}. \tag{48}$$

Then, the PMC ranking satisfies $r^\sigma(y_3) < r^\sigma(y_1) < r^\sigma(y_2)$, as $\mathrm{P}^\sigma(y_3 \succ y_1), \mathrm{P}^\sigma(y_3 \succ y_2), \mathrm{P}^\sigma(y_1 \succ y_2) > 0.5$. However, we have $u_1 = 0.3$ and $u_2 = 0.4$. Since $u_1 < u_2$, it follows that $\Phi^\beta(\sigma)(y_1) < \Phi^\beta(\sigma)(y_2)$ regardless of β, contradicting $r^\sigma(y_1) < r^\sigma(y_2)$. Therefore, Φ^β violates PMC regardless of β.

Next, we show that Φ^∞ satisfies PMC. Consider a profile σ with its PMC ranking r^σ, and let $y^* \in \mathcal{Y}$ be the alternative ranked first in r^σ (i.e., $r^\sigma(y^*) = 1$). Then, y^* must be a Condorcet winner, as $\mathrm{P}^\sigma(y^* \succ y) > 1/2$ for all $y \neq y^*$. Thus, $\pi := \Phi^\infty(\sigma)$ is a deterministic policy with $\pi(y^*) = 1$. Consequently, we have $\pi(y^*) > \pi(y)$ for all $y \neq y^*$ and trivially $\pi(y_i) = \pi(y_j) = 0$ for any $y_i, y_j \neq y^*$, satisfying the condition for PMC. □

Φ^β approximately satisfies PMC as $\beta \to \infty$ if we allow some slack in the rankings (e.g., $\Phi(\sigma)(y_i) \geq \Phi(\sigma)(y_j) - \epsilon$ for some small $\epsilon > 0$) in the definition of PMC. However, exploring this approximate consistency is beyond the scope of this paper and is left for future research.

S Additional details of experiments

S.1 Synthetic dataset

Dataset generation. For the synthetic dataset, we used 10 prompts and 10 responses for the color-preference alignment task, as shown in Table 3. To construct the ground-truth profile σ, we sampled the true (center) rewards independently from the normal distribution $\mathcal{N}(0, 1)$ for each response. We then added i.i.d. random noise from $\mathcal{N}(0, 1)$ to each true reward to generate 1,000 independent rankings. Finally, we drew 10^4 pairwise comparison samples i.i.d. from the true preference function P^σ to train each algorithm.

Evaluation methods. We evaluate the fine-tuned policy using two metrics: (i) win rate against a reference policy (the pretrained model), $\mathbb{E}_{(x,y_1,y_2)\sim(\rho,\pi,\pi_{\mathrm{ref}})}[\mathrm{P}^\sigma(y_1 \succ y_2 \mid x)]$, and (ii) the PPA level $\alpha(\sigma)$. To estimate the fine-tuned policies over responses, we compute the logits of each response and the softmax policy (with temperature 1). We then calculate the win rate and PPA level directly from their definitions using the estimated policy for each prompt. The results are averaged over all prompts.

Table 3: Prompts and responses in synthetic dataset

Prompt (x)	Response (y)
Which color do you find the most appealing?	Red
Which color best represents your personality?	Blue
When decorating your room, what color do you prefer?	Green
What is your favorite color?	Yellow
Which color do you like the most?	Purple
If you had to choose just one color, which would it be?	Orange
Among all colors, what's your top pick?	Pink
If you could only wear one color forever, what would you choose?	Brown
What color makes you feel happiest?	Black
Which color do you prefer most?	White

S.2 Alpaca-GPT4 dataset

Dataset generation. We considered two groups of evaluators, defined across two categories: expertise and style. For the expertise category, evaluators were grouped into two levels: 'elementary school student' and 'PhD student'. For the style category, evaluators were grouped into 'friendly' and 'unfriendly'. The true population distribution was set to $w^\sigma = [0.8, 0.2]$. For each of the 52k instruction prompts from the Alpaca-GPT4 dataset [42], group-specific responses were generated using GPT-4.1 with the prompts listed in Table 4. Then, the pairwise comparison samples are drawn i.i.d. from P^σ.

Table 4: Prompts used for generating responses from each group

Category	Prompt
Expertise	(1) Generate a response that can be easily understood by an elementary school student. (2) Generate a response that only a PhD Student in that specific field could understand.
Style	(1) Generate a response that is friendly, witty, funny, and humorous, like a close friend. (2) Generate a response that answers in an unfriendly manner.

Evaluation methods. To estimate the fine-tuned policies over responses, we sample a response from the policy and use the annotation model (GPT-4.1) to classify its group. Table 5 shows the prompts used to classify the group of generated responses. Based on these classifications, we evaluate the policy's win rate and the PPA level from their definitions.

Table 5: Prompts used for classification

Category	Prompt
Expertise	Does the expertise level of this response align more closely with the elementary level or the PhD student level? Please answer with only one of these exact options: 'elementary' or 'PhD'.
Style	Is this response friendly or unfriendly? Please answer with only one of these exact options: 'friendly' or 'unfriendly'.

S.3 Hyperparameter setting

The Qwen2.5-3B-Instruct model [41] was fine-tuned using each algorithm, where both the reference policy π_{ref} and the data sampling policy π_d were set to the same pretrained model. All algorithms

were trained on the same offline dataset for the same number of iterations. NLHF was not included in the comparison, as the algorithm does not support offline learning. Specific training hyperparameters are provided in Table 6. Each training run utilized one H100 GPU, requiring approximately 0.5–1 hour per epoch with about 20–40GB of memory usage using LoRA.

Table 6: Training hyperparameters

Hyperparameter	Synthetic	Alpaca-GPT4
Training & Reference Model	Qwen2.5-3B-Instruct	Qwen2.5-3B-Instruct
Learning Rate	1e-4	1e-5
Batch Size	8	4
Epochs	3	1
Optimizer	AdamW	AdamW
Gradient Clipping	1.0	1.0
Learning Rate Scheduler	Linear	Linear
Warmup Steps	100	100
KL Coefficient	0.1	0.01
LoRA Rank	32	32
LoRA α	32	32

T Scalable offline algorithm with function approximation

T.1 Offline pairwise comparison dataset

In practical applications of preference learning, the preference function often depends on additional context or state, denoted by $x \in \mathcal{X}$. For instance, in LLMs, x represents the input prompt or conversational history that provides the specific context for generating a preferred response. Accordingly, we define the context-dependent preference function as $\mathrm{P}(\cdot \succ \cdot \mid \cdot) : \mathcal{Y}^2 \times \mathcal{X} \to [0, 1]$, which is unknown and must be estimated from empirical data. We consider an offline dataset of pairwise comparisons $\mathcal{D} = \{(x_i, y_i^w, y_i^l)\}_{i=1}^N$, where y_i^w is preferred over y_i^l under context x_i. Each query is assumed to be drawn i.i.d. from a joint distribution of $\rho(x)$ and $\pi_d(y \mid x)$, and labeled according to the preference function P. Our goal is to use this offline dataset to learn a policy $\pi : \mathcal{X} \mapsto \Delta(\mathcal{Y})$ following the framework introduced in the previous sections.

T.2 Two-Phase offline preference learning algorithm

We approximate a softmax policy proposed in Section 4.3:

$$\pi(y \mid x) := \frac{u(y \mid x) \exp(\beta u(y \mid x))}{\sum_{y \in \mathcal{Y}} u(y \mid x) \exp(\beta u(y \mid x))}, \quad \text{where} \quad u(y \mid x) := \min_{z \in \mathcal{Y}} \mathrm{P}(y \succ z \mid x). \tag{49}$$

Specifically, we use a two-phase algorithm that first estimates u and then estimates π based on u.

Phase 1: Estimating u. To estimate u, we first train the selector model μ using the following loss function, the offline dataset $\mathcal{D}$, and the parameterized function class $\mathcal{F}_\mu$:

$$\hat{\mu} \in \arg\min_{\mu \in \mathcal{F}_\mu} \frac{1}{N} \sum_{i=1}^N \frac{\mu(y_i^l \mid x_i, y_i^w)}{\pi_d(y_i^l \mid x_i)}. \tag{50}$$

Then, the estimated $\hat{u}$ can be obtained from $\hat{u}(y \mid x) = \sum_{z \in \mathcal{Y}} \hat{\mathrm{P}}(y \succ z \mid x) \hat{\mu}(z \mid x, y)$, where $\hat{\mathrm{P}}$ denotes the empirical estimate of the preference function. The derivation of the loss function and the relationship between μ and u are provided in Appendix U.

Phase 2: Estimating π. Let $\hat{\pi}_\beta$ be the normalized softmax policy constructed with $\hat{u}$ following (49). In the second phase, the policy model is trained by minimizing the distance to $\hat{\pi}_\beta$ over a function class $\mathcal{F}_\pi$:

$$\hat{\pi} \in \arg\min_{\pi \in \mathcal{F}_\pi} \mathbb{E}_{x \sim \rho} \Big[L^\pi(\pi(\cdot \mid x), \hat{\pi}_\beta(\cdot \mid x)) \Big]. \tag{51}$$

Here, L^π denotes a divergence or distance metric between two policies. In our experiments, we employ the KL divergence for L^π. Specifically, substituting the KL divergence into L^π from (51), the loss function becomes

$$\begin{aligned}&\mathbb{E}_{x\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot\mid x)\middle\|\sum_{z\in\mathcal{Y}}\hat{P}(\cdot\succ z\mid x)\hat{\mu}(z\mid x,\cdot)\right)\right]\\&=\mathbb{E}_{(x,y)\sim(\rho,\pi_d)}\left[\frac{\pi(y\mid x)}{\pi_d(y\mid x)}\log\frac{\pi(y\mid x)}{\sum_{z\in\mathcal{Y}}\hat{P}(y\succ z\mid x)\hat{\mu}(z\mid x,y)}\right].\end{aligned}\tag{52}$$

Using the offline dataset and function approximation, we obtain

$$\begin{aligned}\hat{\pi}\in\arg\min_{\pi\in\mathcal{F}_\pi}\mathbb{E}_{(x,y^w,y^l)\sim D}\Bigg[&\frac{\pi(y^w\mid x)}{\pi_d(y^w\mid x)}\log\frac{\pi(y^w\mid x)}{(\frac{1}{2}+\frac{1}{2}\hat{\mu}(y^w\mid x,y^l))\exp\left(\frac{\beta}{2}+\frac{\beta}{2}\hat{\mu}(y^w\mid x,y^l)\right)}\\&+\frac{\pi(y^l\mid x)}{\pi_d(y^l\mid x)}\log\frac{\pi(y^l\mid x)}{(\frac{1}{2}-\frac{1}{2}\hat{\mu}(y^l\mid x,y^w))\exp\left(\frac{\beta}{2}-\frac{\beta}{2}\hat{\mu}(y^w\mid x,y^l)\right)}\Bigg].\end{aligned}\tag{53}$$

T.3 Additional techniques for LLM fine-tuning

Regularization via reference policy. Fine-tuning large language models (LLMs) requires maintaining alignment with a reference policy, typically the pretrained model. To prevent excessive drift, we incorporate KL-divergence regularization terms into the training objectives for both the selector model μ and the policy model π. Specifically, we add the following regularization terms to the loss functions in Phase 1 and Phase 2:

$$\beta_\mu\mathbb{E}_{(x,y)\sim(\rho,\pi_d)}\left[D_{\mathrm{KL}}\left(\mu(\cdot\mid x,y)\middle\|\pi_{\mathrm{ref}}(\cdot\mid x,y)\right)\right],\quad\beta_\pi\mathbb{E}_{x\sim\rho}\left[D_{\mathrm{KL}}\left(\pi(\cdot\mid x)\middle\|\pi_{\mathrm{ref}}(\cdot\mid x)\right)\right]\tag{54}$$

Training with single model. To reduce computational cost, we propose to train both μ and π using a single model. This is enabled by encoding structural differences through specialized input formats. Specifically, $\pi(\cdot\mid x)$ selects preferred responses given a prompt, while $\mu(\cdot\mid x,y)$ selects responses given a prompt and a candidate response. By distinguishing these cases with separator tokens, we achieve performance comparable to training separate models, while improving memory usage and training efficiency.

U Derivation of the loss function in phase 1

Step 1: LP reformulation. Recall the definition $u(y\mid x):=\min_{z\in\mathcal{Y}\setminus\{y\}}\mathrm{P}(y\succ z\mid x)$. Each $u(y\mid x)$ can be rewritten as

$$u(y\mid x)=\sum_{z\in\mathcal{Y}}\mathrm{P}(y\succ z\mid x)\mu^*(z\mid x,y),\tag{55}$$

where a *selector distribution* $\mu^*(\cdot\mid x,y)\in\Delta(\mathcal{Y})$ places all its mass on the minimizer of $\mathrm{P}(y\succ\cdot\mid x)$. Such μ^* and the corresponding pointwise minimum can be obtained via the following linear programming (LP):

$$u(y\mid x)=\min_{\mu(\cdot\mid x,y)\in\Delta(\mathcal{Y})}\sum_{z\in\mathcal{Y}}\mathrm{P}(y\succ z\mid x)\mu(z\mid x,y).\tag{56}$$

Step 2: Aggregation of pointwise LPs. Assume the data-generating distribution $\rho(\cdot)$ and $\pi_d(\cdot\mid x)$ have full support. Multiplying and dividing (56) by $\pi_d(z\mid x)$ and then taking the expectation over all $z\in\mathcal{Y}$ gives

$$u(y\mid x)=\min_{\mu(\cdot\mid x,y)\in\Delta(\mathcal{Y})}\mathbb{E}_{z\sim\pi_d(\cdot\mid x)}\left[\frac{\mathrm{P}(y\succ z\mid x)\mu(z\mid x,y)}{\pi_d(z\mid x)}\right].\tag{57}$$

Next, we aggregate these pointwise LPs by multiplying each pointwise objective by $\rho(x)\pi_d(y \mid x)$ and summing over all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We also add the symmetrical term with swapped y and z, which does not change the optimal solution:

$$\mu^* \in \underset{\mu:\mathcal{X}\times\mathcal{Y}\mapsto\Delta(\mathcal{Y})}{\arg\min} \mathbb{E}_{(x,y,z)\sim(\rho,\pi_d,\pi_d)} \left[\frac{\mathrm{P}(y \succ z \mid x)\mu(z \mid x, y)}{\pi_d(z \mid x)} + \frac{\mathrm{P}(z \succ y \mid x)\mu(y \mid x, z)}{\pi_d(y \mid x)} \right]. \quad (58)$$

Step 3: Empirical counterpart. Given an offline preference dataset $\mathcal{D}$, we approximate the expectation in (58) using its empirical counterpart and restrict the function class to a parameterized family $\mathcal{F}_\mu$:

$$\hat{\mu} \in \underset{\mu\in\mathcal{F}_\mu}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \frac{\mu(y_i^l \mid x_i, y_i^w)}{\pi_d(y_i^l \mid x_i)}. \quad (59)$$

Given the estimate $\hat{\mu}$, we can estimate $\hat{u}$ using (56) with the estimated preference function $\hat{\mathrm{P}}$:

$$\hat{\mathrm{P}}(y \succ z \mid x) := \begin{cases} \frac{N(x,y,z)}{N(x,y,z)+N(x,z,y)} & \text{if } N(x, y, z) + N(x, z, y) > 0, \\ 1/2, & \text{otherwise}, \end{cases} \quad (60)$$

where $N(x, y, z) := |\{i \in [N] \mid (x_i, y_i^w, y_i^l) = (x, y, z)\}|$.