SharedRep-RLHF: A Shared Representation Approach to RLHF with Diverse Preferences

Arpan Mukherjee[1*], **Marcello Bullo**[1*], **Deniz Gündüz**[1]
[1]Imperial College London
{a.mukherjee, m.bullo21, d.gunduz}@imperial.ac.uk

Abstract

Uniform-reward reinforcement learning from human feedback (RLHF), which trains a single reward model to represent the preferences of all annotators, fails to capture the diversity of opinions across sub-populations, inadvertently favoring dominant groups. The state-of-the-art, MaxMin-RLHF, addresses this by learning group-specific reward models, and by optimizing for the group receiving the minimum reward, thereby promoting fairness. However, we identify that a key limitation of MaxMin-RLHF is its poor performance when the minimum-reward group is a minority. To mitigate this drawback, we introduce a novel framework termed *SharedRep-RLHF*. At its core, SharedRep-RLHF learns and leverages *shared preference traits* in annotations among various groups, in contrast to learning separate reward models across groups. We first show that MaxMin-RLHF is provably suboptimal in learning shared traits, and then quantify the sample complexity of SharedRep-RLHF. Experiments across diverse natural language tasks showcase the effectiveness of SharedRep-RLHF compared to MaxMin-RLHF with a gain of up to 20% in win rate.

Code — https://github.com/marcellobullo/sharedrep-rlhf
Extended version — https://arxiv.org/abs/2509.03672

1 Motivation & Overview

The success of large language models (LLMs) is largely attributed to their ability to generate responses which adhere to human values and behavior. The art of steering LLMs to elicit human-aligned responses is known as *alignment*, and *reinforcement learning from human feedback* (RLHF) serves as the cornerstone for aligning LLMs (Christian 2021; Ouyang et al. 2022; Bai et al. 2022). The canonical RLHF pipeline is a three-step process consisting of (i) supervised fine-tuning (SFT), (ii) reward-modeling, and (iii) reinforcement learning (RL) fine-tuning (Wang et al. 2023). In this paper, our focus is on (ii) reward-modeling and (iii) RL fine-tuning. Reward modeling involves fitting a reward model to a *preference dataset*, and RL fine-tuning uses policy-gradient methods (e.g., proximal policy optimization (PPO) (Schulman et al. 2017) and group relative preference optimization (GRPO) (Shao et al. 2024)) to obtain a policy that maximizes a (regularized) average reward objective based on the modeled reward in the second step.

Uniform versus diverse preference. In this paper, we focus on *offline RLHF*. Existing investigations on RLHF (Ouyang et al. 2022; Bai et al. 2022; Stiennon et al. 2022; Liu et al. 2024; Christian 2021; Xie et al. 2024; Pang et al. 2024; Cen et al. 2024; Zhang et al. 2024; Casper et al. 2023) mostly focus on a uniform reward model representing all annotators. This is indicative of a monolithic view of the world, where annotators are assumed to agree on *all* aspects, ignoring a vast array of diverse traits intrinsic to annotators from distinct subpopulations. On the contrary, a holistic view of the world captures both subtle and profound nuances in human traits due to individual differences, such as cognitive styles and biases, cultural and societal factors such as language and dialect, ethical and moral values, as well as task-specific differences such as domain expertise and context dependence. In one of the early investigations into the possibility of misalignment of the views of LLMs (Santurkar et al. 2023), it was concluded that there was *significant* misalignment across different demographics, and in many cases, LLMs' views were considerably biased towards certain groups. In order to promote group fairness, Chakraborty et al. introduced *MaxMin-RLHF*, a framework to steer the LLM policy to account for the preferences of the worst-case (potentially the minority) group. In this framework, group-specific reward models are learnt from the preference dataset, and the canonical KL-regularized objective for the RL fine-tuning step is replaced by a MaxMin objective, trying to maximize the (regularized) reward for the *worst* (lowest reward) group. This framework has been shown to promote fairness with respect to the preferences across diverse groups, and has since been widely investigated for addressing diversity in human preferences from a fairness viewpoint (Ramesh et al. 2024; Son et al. 2025). A closely related research direction is that of personalized RLHF (Jang et al. 2023; Chen et al. 2024; Poddar et al. 2024; Dong et al. 2025), where the goal is to adapt the LLMs to the preferences of *individual* users. Note that we do not aim to solve personalized RLHF, i.e., personalizing to preferences per user. Rather, our focus is on promoting group-fairness in handling distinct preferences across multiple groups.

[*]These authors contributed equally.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

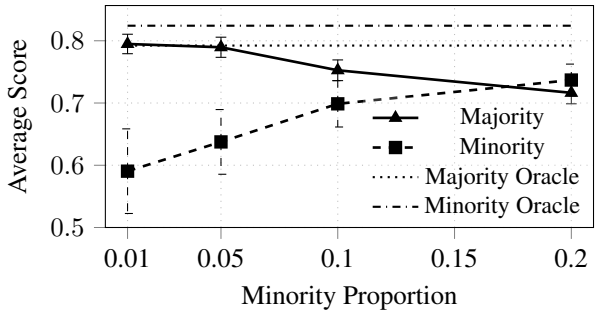

Figure 1: Average group scores versus minority proportions of MaxMin-RLHF on controlled sentiment analysis for the IMDb dataset (Maas et al. 2011).

Drawbacks of MaxMin-RLHF. The MaxMin objective, in principle, promotes fairness by steering the policy to prioritize the group with the lowest expected reward (reward minority). However, what happens when reward minority is also the *underrepresented group*, i.e., the group with the fewest annotations in the dataset? Does MaxMin-RLHF preserve its group-fairness property when such low-prevalence group becomes increasingly small? In Figure 1, we replicate the "small scale experiment" setting in (Chakraborty et al. 2024) for sentiment analysis using the IMDb dataset (Maas et al. 2011), in which the population is split into a *majority* and a *minority* group. The majority prefers positive reviews, while the minority prefers concise reviews. As the proportion of the minority decreases, Figure 1 shows how MaxMin-RLHF alignment generates responses with worse scores for the minority, exhibiting a *significant gap* with the oracle policy, which assumes to know the ground truth reward model. MaxMin-RLHF requires a minimal level of minority share ($\approx 20\%$) before it can reach some form of parity with the majority. These effects demonstrate that MaxMin-RLHF is sample-inefficient, suboptimal for both subgroups, and poorly suited to skewed data settings.

Contributions. We address these limitations of the MaxMin-RLHF framework in settings where minority annotators are underrepresented. Our key insight is that, despite group-specific nuances, preference data across groups often reflect common underlying human values. We propose to learn these *shared values* from the entire dataset, while capturing group-specific nuances using annotations from each group. Our contributions are threefold.

Framework: We introduce a novel framework, *shared representation RLHF* (SharedRep-RLHF), that addresses the estimation inaccuracies of MaxMin-RLHF by learning and leveraging shared traits among groups. The goal, in this premise, is to solve the MaxMin objective to produce a policy that promotes fairness across groups in its generated responses. SharedRep-RLHF is a generalization of the MaxMin-RLHF framework, which is recovered in the special case when groups do not share any common trait. Furthermore, we propose an algorithm called *SharedRep-RLHF* that provably optimizes the MaxMin objective proposed in (Chakraborty et al. 2024).

Analytical contributions: We statistically characterize various algorithmic facets of SharedRep-RLHF. We make three analytical contributions. First, we derive confidence sets for maximum likelihood estimates (MLEs) in the SharedRep-RLHF framework, which scale inversely with the square root of the size of the *entire dataset* (Lemma 2), as opposed to the size of group-specific datasets, which has been reported in (Chakraborty et al. 2024). Second, building on this result, we show that SharedRep-RLHF offers provably better estimation fidelity compared to MaxMin-RLHF (Theorem 1). Third, we characterize the sample complexity of SharedRep-RLHF in the probabably approximately correct (PAC) framework in Theorem 2. We find that the additional number of samples required to enforce the MaxMin objective over the canonical regularized reward objective (Xiong et al. 2024) is of the order $O\big(1/\Delta_{\min}^4\big)$, where $\Delta_{\min}$ is the minimum gap between conditional entropies of the induced Gibbs distributions of the groups. In the process, we also show that the group that has the *minimum reward* is equivalent to the group exhibiting *maximal entropy* in its induced Gibbs distribution (Lemma 3). Note that the framework in (Chakraborty et al. 2024) does not come with statistical guarantees; hence, we provide the first result on the sample complexity of MaxMin-RLHF.

Experiments: We empirically validate SharedRep-RLHF on three diverse tasks – *controlled sentiment analysis*, *mathematical reasoning*, and *single turn dialogue* – under varying degrees of minority underrepresentation. SharedRep-RLHF consistently outperforms MaxMin-RLHF in both mean minority score and win rate, especially in low minority proportion regimes, demonstrating its effectiveness and robustness in aligning the underlying LLM with underrepresented group preferences leveraging shared values across group prefernces.

2 Preliminaries

In this section, we provide a brief background on the MaxMin-RLHF pipeline (Chakraborty et al. 2024) and subsequently introduce our proposed SharedRep-RLHF framework, a generalization of MaxMin-RLHF that is cognizant of shared preferences among the groups.

MaxMin-RLHF. The premise in RLHF is that the learner has access to a preference dataset denoted by $\mathcal{D} \triangleq \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{y}_i', z_i)\}_{i=1}^N$. Here, for each prompt $\mathbf{x}_i$, $i \in [N]$, $\mathbf{y}_i, \mathbf{y}_i' \sim \pi_{\text{ref}}(\cdot \mid \mathbf{x}_i)$ denote two responses generated by the reference model conditioned on prompt $\mathbf{x}_i$, and $z_i \in \{0, 1\}$ denotes the annotators' preference, where $z_i = 1$ if the annotator prefers $\mathbf{y}_i$ as a response to prompt $\mathbf{x}_i$, and $z_i = 0$ otherwise. We denote the space of prompts by $\mathcal{X}$ and the space of responses by $\mathcal{Y}$. To capture the diversity in human preferences, Chakraborty et al. cluster the annotator samples into U groups and interpret these clusters as reflecting underlying human subpopulations. Let $\mathcal{H}$ denote the set of all annotators. Accordingly, we assume that $\mathcal{H} = \cup_{u=1}^U \mathcal{H}_u$ such that $\mathcal{H}_u \cap \mathcal{H}_v = \emptyset$ for all $u \neq v$. Based on the partitioning of the annotators, Chakraborty et al. assume that each subpopulation $u \in [U]$ has an intrinsic parameter $\boldsymbol{\theta}_u^\star$ that models its preferences using the Bradley-Terry (BT)

model (Bradley and Terry 1952), i.e.,

$$\mathbb{P}_u\Big(z = 1 \mid \mathbf{x}, \mathbf{y}, \mathbf{y}'\Big) \triangleq \sigma\Big(r_{\boldsymbol{\theta}_u^\star}(\mathbf{x}, \mathbf{y}) - r_{\boldsymbol{\theta}_u^\star}(\mathbf{x}, \mathbf{y}')\Big), \quad (1)$$

where σ denotes the logistic function, $\mathbb{P}_u$ denotes the preference measure corresponding to the subpopulation $u \in [U]$ and $r_{\boldsymbol{\theta}_u^\star}$ denotes a parametric representation of its intrinsic reward function $r_u^\star$ guiding the preferences. We further assume that the partitioning of the annotators is *known*, i.e., for each annotator $i \in \mathcal{H}$, we know the subpopulation to which they belong, i.e., $i \in \mathcal{H}_u$[1], $u \in [U]$. In this setting, distinctly to uniform preference RLHF, we can form estimates of group-specific parameters $\widehat{\boldsymbol{\theta}}_u$ for every $u \in [U]$ by performing reward-modeling using group-specific preference data. Once we form these estimates, Chakraborty et al. propose to solve the MaxMin reward objective:

$$\pi_{\text{MaxMin}} \in \arg\max_{\pi} \Bigg\{ \min_{u \in [U]} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi(\cdot \mid \mathbf{x})} \Big[r_{\widehat{\boldsymbol{\theta}}_u}(\mathbf{x}, \mathbf{y}) - \beta D_{\mathsf{KL}}\big(\pi(\cdot \mid \mathbf{x}) \| \pi_{\text{ref}}(\cdot \mid \mathbf{x})\big) \Big] \Bigg\}, \quad (2)$$

inspired by the egalitarian principle in social choice theory (Sen 2017). The objective corresponds to maximizing the regularized value function corresponding to the *minimum-reward* subpopulation.

SharedRep-RLHF. Chakraborty et al. treat subpopulation-specific preferences independently, as their method ignores potential correlations or shared structure across groups. In practice, however, some prompts and responses may elicit similar preferences across subpopulations, and exploiting such shared latent structure can improve sample efficiency. Specifically, there exist common preference traits across the whole population $\mathcal{H}$ that can be estimated from the preference data across all subpopulations. We propose to view the problem from the lens of representation learning in bandits (Yang et al. 2020). Specifically, we assume that there is a universal feature extractor matrix $\mathbf{B}^\star \in \mathbb{R}^{d\times K}$ across subpopulations, whose columns may be interpreted as representing various *shared latent preference features* (usually $K \ll d$). For each subpopulation $u \in [U]$, the intrinsic parameters may be expressed as $\boldsymbol{\theta}_u^\star \triangleq \mathbf{B}^\star \mathbf{w}_u^\star$, where $\mathbf{w}_u^\star \in \Delta^{K-1}$ is a mixing coefficient in the probability simplex of order K. We call this framework *shared representation RLHF* (SharedRep-RLHF) in light of the common feature extractor $\mathbf{B}^\star$. In our formulation, for those traits that are shared among the entire population, the mixing coefficients (or weights) would be similar for all subpopulations. Subpopulation-specific traits, on the other hand, would exhibit disparity in their weights learned across various subpopulations. Note that this framework is complementary to (Conitzer et al. 2024), which models group-specific preferences from elicited attributes. By instead learning shared latent features directly from preference data, our approach offers greater flexibility than (Chakraborty et al. 2024). To elaborate, when we have fewer preference data for a minority subpopulation, attempting to estimate the corresponding intrinsic parameter from limited data points might result in extremely inaccurate or unreliable reward estimates. In the proposed SharedRep-RLHF framework, data from the entire population can be used to estimate the feature extractor. Hence, we hope to achieve better estimation accuracy leveraging the feature extractor compared to the framework in (Chakraborty et al. 2024). In order to statistically analyze the performance of the proposed algorithm, we make the following assumptions which are commonly adopted for the analysis of RLHF algorithms.

Assumption 1 (Linear reward)**.** *We assume that the reward function is linearly parameterized, i.e., for each subpopulation $u \in [U]$, we have $r_{\mathbf{B}\mathbf{w}_u}(\mathbf{x}, \mathbf{y}) \triangleq \big\langle \phi(\mathbf{x}, \mathbf{y}), \mathbf{B}\mathbf{w}_u \big\rangle$ for any feature extractor $\mathbf{B} \in \mathcal{B}$ and $\mathbf{w}_u \in \Delta^{K-1}$, such that $\mathcal{B} \triangleq \big\{\mathbf{B} \in \mathbb{R}^{d\times K} : \|\mathbf{B}_{:,k}\|_2 \le B_{\max}\ ,\ \forall\ k \in [K]\big\}$, $\mathbf{B}_{:,k}$ denotes the k^{th} column of matrix $\mathbf{B}$, and $\phi(\mathbf{x}, \mathbf{y})$ denotes a known embedding of the concatenated (prompt, response) pair $(\mathbf{x}, \mathbf{y})$. Furthermore, we assume that $\|\phi(\cdot, \cdot)\| \le L_{\max}$.*

Assumption 1 has been extensively used for deriving statistical guarantees in the RLHF literature (Kong and Yang 2022; Saha, Pacchiano, and Lee 2023; Zhu, Jordan, and Jiao 2023; Xiong et al. 2024; Foster, Mhammedi, and Rohatgi 2025). While (1) identifies each group's reward only up to a group-specific additive constant, Assumption 1 restricts the class enough to make cross-group comparisons well defined. In experiments, ϕ is generally chosen as the final layer of the frozen SFT model, after removing the logits and the softmax layers. In contrast to the existing norm of adding a single neuron after the last layer of the frozen SFT backbone with fully connected and trainable weights, the SharedRep-RLHF framework adds a shared linear layer $\mathbf{B}$, which captures the shared preferences between subpopulations, followed by a fully connected neuron for each subpopulation.

Assumption 2 (Reward gap)**.** *For any subpopulation $u \in [U]$, and for any pair of parameters in the hypothesis classes $\mathcal{B}$ and Δ^{K-1}, there exists a prompt for which the minimum reward difference between the chosen and rejected responses is bounded away from 0. Specifically, we assume that $\xi_u > 0$ for all $u \in [U]$, where we have defined*

$$\xi_u \triangleq \inf_{\substack{\mathbf{B}\in\mathcal{B} \\ \mathbf{w}_u \in \Delta^{K-1}}} \max_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{y}\neq\mathbf{y}'} \Big| \big\langle \phi(\mathbf{x}, \mathbf{y}) - \phi(\mathbf{x}, \mathbf{y}'), \mathbf{B}\mathbf{w}_u \big\rangle \Big|.$$

Assumption 2 implies that for *every* model in the parameter space, there always exist a prompt which would yield a minimum reward difference that is bounded away from 0. This is a very natural assumption, as otherwise we would have the possibility of models under which it is *impossible to differentiate* between distinct pairs of responses. In the reward-modeling stage, we form estimates $\widehat{\mathbf{B}}$ for the feature extractor $\mathbf{B}^\star$, and $\widehat{\mathbf{W}} \triangleq [\widehat{\mathbf{w}}_1, \cdots, \widehat{\mathbf{w}}_U]$ corresponding to the weights $\mathbf{W}^\star \triangleq [\mathbf{w}_1^\star, \cdots, \mathbf{w}_U^\star]$. Finally, we assume that there are no degenerate states.

Assumption 3 (Non-degeneracy)**.** *The distribution over prompts, denoted by ρ, satisfies $\rho_{\min} > 0$, where we have defined $\rho_{\min} \triangleq \min_{\mathbf{x}\in\mathcal{X}} \rho(\mathbf{x})$.*

[1]Alternatively, we can always invoke (Chakraborty et al. 2024, Algorithm 1) to learn an appropriate clustering.

Next, similar to MaxMin-RLHF, our goal is to design a policy that maximizes the KL-regularized value function comprising the estimated reward models. Specifically, the learner's goal is to design a policy that maximizes the KL-regularized MaxMin value function, i.e.,

$$\pi^\star \in \arg\max_\pi J_{\mathrm{MaxMin}}(\pi) \triangleq \min_{u\in[U]} \Big\{ \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi(\cdot\mid\mathbf{x})} \Big[r_{\boldsymbol{\theta}_u^\star}(\mathbf{x},\mathbf{y}) - \beta D_{\mathsf{KL}}\big(\pi(\cdot\mid\mathbf{x})\|\pi_{\mathrm{ref}}(\cdot\mid\mathbf{x})\big)\Big]\Big\}. \tag{3}$$

3 Estimator & Performance Comparison

In this section, we propose an estimator for the reward-modeling stage, which is subsequently used to optimize the MaxMin objective. Furthermore, we asses the performance of the proposed estimator, comparing it to the MaxMin-RLHF framework proposed in (Chakraborty et al. 2024), and showcasing a provable gain over the MaxMin framework. All proofs are deferred to the appendix for brevity.

Definitions. Let us first introduce a few notations. For specifying the estimator, for any data point $i \in [N]$, let us define the binary cross-entropy loss

$$\ell_i(\mathbf{B},\mathbf{w}_u) \triangleq z_i \log\Big(\sigma\Big(r_{\mathbf{B}\mathbf{w}_u}(\mathbf{x}_i,\mathbf{y}_i) - r_{\mathbf{B}\mathbf{w}_u}(\mathbf{x}_i,\mathbf{y}_i')\Big)\Big) + (1-z_i)\log\Big(1-\sigma\Big(r_{\mathbf{B}\mathbf{w}_u}(\mathbf{x}_i,\mathbf{y}_i) - r_{\mathbf{B}\mathbf{w}_u}(\mathbf{x}_i,\mathbf{y}_i')\Big)\Big).$$

The MLE is obtained by minimizing the aggregate loss, i.e.,

$$\widehat{\mathbf{B}},\widehat{\mathbf{W}} \in \mathop{\arg\min}_{\mathbf{B}\in\mathcal{B},\mathbf{w}_u\in\Delta^{K-1},\forall u\in[U]} \sum_{i\in[N]} \ell_i(\mathbf{B},\mathbf{w}_u). \tag{4}$$

In order to capture the estimation fidelity based on the offline dataset $\mathcal{D}$, following prior art (Zhu, Jordan, and Jiao 2023; Das et al. 2024), we adopt the *unregularized* value function as the performance metric. Specifically, for any policy π, and for any subpopulation $u \in [U]$, let us define the unregularized value function under the ground truth as

$$J_u(\pi) \triangleq \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi(\cdot\mid\mathbf{x})}\Big[r_{\boldsymbol{\theta}_u^\star}(\mathbf{x},\mathbf{y})\Big]. \tag{5}$$

Furthermore, let $\pi_u^\star \in \arg\max_\pi J_u(\pi)$ denote the optimal policy maximizing the unregularized value function for subpopulation $u \in [U]$. Let $\mathfrak{A}$ denote an algorithm used for forming estimates of the model parameters. The fidelity of estimation is captured through the value function gap due to the induced policy $\widehat{\pi}_{\mathfrak{A}}$, i.e., for any group $u \in [U]$,

$$\mathrm{SubOpt}_u(\widehat{\pi}_{\mathfrak{A}}) \triangleq J_u(\pi_u^\star) - J_u(\widehat{\pi}_{\mathfrak{A}}). \tag{6}$$

Pessimism in offline RLHF. It is well-established in the offline RLHF literature (Zhu, Jordan, and Jiao 2023; Xiong et al. 2024) that using MLEs (i.e., $\mathfrak{A} = \mathrm{MLE}$) may result in a potentially unbounded suboptimality gap. A mechanism to handle this issue is the principle of *pessimism*. Using $\mathfrak{A} = \mathrm{MM}$ and $\mathfrak{A} = \mathrm{SR}$ to denote the MaxMin-RLHF and the SharedRep-RLHF frameworks, and setting $N_u = |\mathcal{H}_u|$, we have the following results.

Lemma 1 (Zhu, Jordan, and Jiao (2023))**.** *Recall that $\widehat{\boldsymbol{\theta}}_u$ denotes the MLE formed under the MaxMin-RLHF framework from the subpopulation-specific data for each $u \in [U]$. Under Assumption 1, for any $\lambda \in \mathbb{R}_+$ and $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\big\|\widehat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_u^\star\big\|_{\Sigma_u+\lambda\mathbb{I}} \le C_{\mathrm{MM}}\sqrt{\frac{1}{N_u}C_\delta + \lambda B_{\max}^2}, \tag{7}$$

where $\gamma \triangleq 1/\big(2+\exp(-L_{\max}B_{\max})+\exp(L_{\max}B_{\max})\big)$, $\Sigma_u \triangleq (1/N_u)\sum_{i\in\mathcal{H}_u}\big(\phi(\mathbf{x}_i,\mathbf{y}_i)-\phi(\mathbf{x}_i,\mathbf{y}_i')\big)\big(\phi(\mathbf{x}_i,\mathbf{y}_i)-\phi(\mathbf{x}_i,\mathbf{y}_i')\big)^\top$, $C_\delta \triangleq \frac{d+\log(1/\delta)}{\gamma^2}$, , and $C_{\mathrm{MM}} \in \mathbb{R}_+$ is a universal constant.

Remark 1. *We observe from Lemma 1 that the accuracy of subpopulation-specific reward estimates rely heavily on the size of the subpopulation N_u. Specifically, if N_u is small, the estimation error $\|\widehat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}_u^\star\|_{\Sigma_u+\lambda\mathbb{I}}$ is only controlled in the few directions covered by Σ_u, and the confidence bound is relatively large. This matches our experimental results in Figure 1: the smaller the subpopulation size, the worse the estimation accuracy, and consequently, the worse the performance of the MaxMin-RLHF framework.*

Unlike the MaxMin-RLHF framework, we propose to learn traits shared among subpopulations, and as a result, improve the estimation accuracy over MaxMin-RLHF. The SharedRep-RLHF framework learns a shared feature extractor $\mathbf{B}$ from the entire dataset $\mathcal{D}$. This aids in the estimation of $\boldsymbol{\theta}_u^\star$ for every $u \in [U]$, since we are no longer restricted by directions which are exclusively covered by Σ_u. More specifically, we have the following concentration on the MLEs in the SharedRep-RLHF framework.

Lemma 2 (Estimator concentration)**.** *Under Assumption 1, for any $\lambda \in \mathbb{R}_+$ and $\delta \in (0,1)$, with probability at least $1-\delta$, for every subpopulation $u \in [U]$ we have*

$$\big\|\widehat{\mathbf{B}}\widehat{\mathbf{w}}_u - \mathbf{B}^\star\mathbf{w}_u^\star\big\|_{\Sigma+\lambda\mathbb{I}} \le C_{\mathrm{SR}}\sqrt{\frac{1}{N}C_\delta + \lambda B_{\max}^2}\,, \tag{8}$$

where $\Sigma \triangleq \frac{1}{N}\sum_{i\in[N]}\big(\phi(\mathbf{x}_i,\mathbf{y}_i)-\phi(\mathbf{x}_i,\mathbf{y}_i')\big)\big(\phi(\mathbf{x}_i,\mathbf{y}_i)-\phi(\mathbf{x}_i,\mathbf{y}_i')\big)^\top$, and $C_{\mathrm{SR}} \in \mathbb{R}_+$ is a universal constant.

Pessimistic value functions. Based on Lemmas 1 and 2, we now define confidence sets, which indicate the estimation fidelity in the MaxMin and SharedRep frameworks, respectively, as follows. For every $u \in [U]$,

$$\Theta_{\mathrm{MM}}(\widehat{\boldsymbol{\theta}}_u) \triangleq \Big\{\boldsymbol{\theta}\in\Theta : \|\widehat{\boldsymbol{\theta}}_u - \boldsymbol{\theta}\|_{\Sigma_u+\lambda\mathbb{I}} \le C_{\mathrm{MM}}\sqrt{\frac{1}{N_u}C_\delta+\lambda B_{\max}^2}\Big\}, \tag{9}$$

$$\Theta_{\mathrm{SR}}(\widehat{\mathbf{B}},\widehat{\mathbf{w}}_u) \triangleq \Big\{(\mathbf{B},\mathbf{w})\in\mathcal{B}\times\Delta^{K-1} : \|\widehat{\mathbf{B}}\widehat{\mathbf{w}}_u - \mathbf{B}\mathbf{w}\|_{\Sigma+\lambda\mathbb{I}} \le C_{\mathrm{SR}}\sqrt{\frac{1}{N}C_\delta+\lambda B_{\max}^2}\Big\}. \tag{10}$$

Furthermore, we define *pessimistic value functions* based on the above confidence sequences in each framework for

any policy π. The pessimism comes from the parameter estimate, which is chosen as the one that yields the *minimal value function* within the confidence sets. Formally, in the MaxMin-RLHF framework, for a policy π and for any subpopulation $u \in [U]$, we define the estimated pessimistic value function and its optimal policy as

$$\widehat{J}_u^{\mathrm{MM}}(\pi) \triangleq \min_{\boldsymbol{\theta} \in \Theta_{\mathrm{MM}}} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi(\cdot \mid \mathbf{x})} \Big[\big\langle \phi(\mathbf{x}, \mathbf{y}), \boldsymbol{\theta} \big\rangle \Big]$$

and its optimal policy as $\widehat{\pi}_u^{\mathrm{MM}} \in \arg\max_\pi \widehat{J}_u(\pi)$. Similarly, in the SharedRep-RLHF framework, for any $u \in [U]$, we define the estimated pessimistic value function as

$$\widehat{J}_u^{\mathrm{SR}}(\pi) \triangleq \min_{(\mathbf{B},\mathbf{w}) \in \Theta_{\mathrm{SR}}} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi(\cdot \mid \mathbf{x})} \Big[\big\langle \phi(\mathbf{x}, \mathbf{y}), \mathbf{B}\mathbf{w} \big\rangle \Big],$$

and its optimal policy as $\widehat{\pi}_u^{\mathrm{SR}} \in \arg\max_\pi \widehat{J}_u^{\mathrm{SR}}(\pi)$. Note that prior work (Zhu, Jordan, and Jiao 2023) proposed defining value functions shifted by a fixed bias vector $\boldsymbol{\nu}$. Our results remain valid under this formulation; however, for clarity of presentation, we omit the bias vector in our definitions.

Performance comparison. Having defined the estimated value functions $\widehat{J}_u^{\mathrm{MM}}$ and $\widehat{J}_u^{\mathrm{SR}}$, we now demonstrate that the estimation framework of SharedRep-RLHF offers provably improved performance over MaxMin-RLHF, using the pessimistic value function as the performance metric.

Theorem 1 (Performance comparison)**.** *Under Assumptions 1,2, and 3, if $\widehat{\pi}_u^{\mathrm{MM}} \neq \widehat{\pi}_u^{\mathrm{SR}}$, for any $u \in [U]$ and for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\begin{aligned} &\mathrm{SubOpt}_u(\widehat{\pi}_u^{\mathrm{MM}}) - \mathrm{SubOpt}_u(\widehat{\pi}_u^{\mathrm{SR}}) \\ &\geq \rho_{\min}\xi_u - 2\eta^{\mathrm{SR}}(N,\lambda,\delta) \cdot \mathbb{E}_{\mathbf{x}\sim\rho}\left[\kappa^{\mathrm{MM}}(\Sigma,\lambda,\mathbf{x})\right], \end{aligned} \tag{11}$$

with $\kappa^{\mathrm{MM}}(\Sigma,\lambda,\mathbf{x}) \triangleq \left\| \mathbb{E}_{\mathbf{y}\sim\widehat{\pi}_u^{\mathrm{MM}}(\cdot \mid \mathbf{x})}\left[\phi(\mathbf{x},\mathbf{y})\right] \right\|_{(\Sigma+\lambda\mathbb{I})^{-1}}$, and $\eta^{\mathrm{SR}}(N,\lambda,\delta) \triangleq C_{\mathrm{SR}}\sqrt{\frac{1}{N}C_\delta + \lambda B_{\max}^2}$.

Theorem 1 shows that when $\lambda = 1/N$, as the dataset size grows, the performance gain of SharedRep-RLHF estimation becomes evident. Asymptotically, the performance gap is quantified through a constant term ξ_u for any subpopulation $u \in [U]$, which captures the (prompt-wise) maximal minimum reward gap between the chosen and rejected responses that exists in the dataset. Additionally, the term $\kappa^{\mathrm{MM}}(\Sigma,\lambda,\mathbf{x})$ can be upper-bounded by $L_{\max}\sqrt{\lambda_{\min}(\Sigma+\lambda\mathbb{I})}$, where $\lambda_{\min}(\mathbf{A})$ denotes the smallest eigenvalue of matrix $\mathbf{A}$. Consequently, this multiplicative factor is, in the worst-case, $O(1)$.

4 Sample Complexity of SharedRep-RLHF

In canonical RLHF, GRPO and PPO are used to find a near-optimal solution to (3), assuming that $U = 1$. In this Section, we extend this by proposing an algorithm tailored to optimize (3) across multiple subpopulations. After the reward learning phase, the algorithm applies GRPO (or PPO) using the value function associated with the subpopulation exhibiting the lowest average reward, similarly to MaxMin-RLHF. Furthermore, contrary to Chakraborty et al., we provide a sample complexity analysis of the proposed method, establishing conditions under which it achieves an ε-accurate approximation of the optimal value function.

SharedRep-RLHF Algorithm. We propose the SharedRep-RLHF algorithm which is grounded in the MaxMin-RLHF framework. We form MLEs of $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{W}}$, which subsequently serve to construct the corresponding confidence sets via Lemma 2. After reward-modeling, we leverage the MaxMin objective to align the SFT model based on pessimistic reward estimates from the previous step. To implement pessimism, we follow (Xiong et al. 2024), which proposes to subtract an uncertainty term $\Gamma(\Sigma,\lambda,\pi)$ scaled by the confidence width $\eta^{\mathrm{SR}}(N,\lambda,\delta)$ from the average reward under a policy π. Finally, we adopt the MaxMin objective to find a policy that maximizes the worst-case pessimistic KL-regularized reward, i.e.,

$$\begin{aligned} \widetilde{\pi}^{\mathrm{SR}} \in \arg\max_\pi \min_{u\in[U]} \Big\{ &\mathbb{E}_{\substack{\mathbf{x}\sim\rho \\ \mathbf{y}\sim\pi(\cdot\mid\mathbf{x})}} \Big[r_{\widehat{\mathbf{B}}\widehat{\mathbf{w}}_u}(\mathbf{x},\mathbf{y}) \Big] \\ &- \eta^{\mathrm{SR}}(N,\lambda,\delta)\,\Gamma(\Sigma,\lambda,\pi) - \beta D_{\mathsf{KL}}\left(\pi(\cdot \mid \mathbf{x}) \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})\right) \Big\}. \end{aligned} \tag{12}$$

Performance Guarantees. The SharedRep-RLHF algorithm involves finding a policy π that maximizes the estimated average reward corresponding to the *worst-case subpopulation*, i.e.,

$$\widehat{u} \in \arg\min_{u\in[U]} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\widetilde{\pi}^{\mathrm{SR}}(\cdot \mid \mathbf{x})} \Big[r_{\widehat{\mathbf{B}}\widehat{\mathbf{w}}_u}(\mathbf{x},\mathbf{y}) \Big]. \tag{13}$$

In order to provide a sample complexity analysis of the SharedRep-RLHF algorithm, we first make an observation that the subpopulation corresponding to the worst-case reward, as stated in (13), can equivalently be expressed as the subpopulation that maximizes the entropy of the intrinsic conditional Gibbs distribution. For $\{r_u : u \in [U]\}$ let us define

$$\begin{aligned} \pi_r \in \arg\max_\pi \min_{u\in[U]} \Big\{ &\mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi(\cdot \mid \mathbf{x})} \Big[r_u(\mathbf{x},\mathbf{y}) \Big] \\ &- \beta D_{\mathsf{KL}}\big(\pi(\cdot \mid \mathbf{x}) \| \pi_{\mathrm{ref}}(\cdot \mid \mathbf{x})\big) \Big\}. \end{aligned} \tag{14}$$

Furthermore, let u_r denote the corresponding worst-case subpopulation, i.e.,

$$u_r \in \arg\min_{u\in[U]} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_r(\cdot \mid \mathbf{x})} \Big[r_u(\mathbf{x},\mathbf{y}) \Big]. \tag{15}$$

Let us define the Gibbs distributions induced by a reward function r given a prompt $\mathbf{x}$ as

$$\nu_r(\mathbf{y} \mid \mathbf{x}) \triangleq \frac{\pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) \exp\left(\frac{1}{\beta} r(\mathbf{x},\mathbf{y})\right)}{\sum_{\mathbf{y}} \pi_{\mathrm{ref}}(\mathbf{y} \mid \mathbf{x}) \exp\left(\frac{1}{\beta} r(\mathbf{x},\mathbf{y})\right)}, \tag{16}$$

and let $H(\nu_r(\cdot \mid \mathbf{X}))$ denote the conditional entropy of the Gibbs distribution corresponding to the subpopulation $u \in [U]$, i.e., $H(\nu_r(\cdot \mid \mathbf{X})) \triangleq \mathbb{E}_{\mathbf{x}\sim\rho}[H(\nu_r(\cdot \mid \mathbf{x}))]$. In the following lemma, we establish the equivalence between selecting the subpopulation that exhibits the minimal average reward, and selecting the one that maximizes the (conditional) entropy of its Gibbs distribution.

Lemma 3 (Worst-case subpopulation)**.** *The worst-case subpopulation u_r, obtained by solving the MaxMin objective over the collection $\{r_u : u \in [U]\}$ in* (15)*, is equivalently characterized as the subpopulation attaining the maximal conditional entropy of the associated Gibbs distribution, i.e., $u_r \in \arg\max_{u\in[U]} H(\nu_{r_u}(\cdot \mid \mathbf{X}))$.*

We leverage Lemma 3 to derive an upper bound on the sample complexity of the SharedRep-RLHF algorithm. Specifically, we will quantify a *sufficient* number of samples N that ensures an (ε, δ)-PAC guarantee on the estimated regularized MaxMin value function. In order to specify this, we introduce a few notations. For any $u \in [U]$, let $\nu^\star_u(\cdot \mid \mathbf{x})$ denote the Gibbs distribution induced by the true reward model $r_{\boldsymbol{\theta}^\star_u}$ as specified in (16). Furthermore, let $u^\star$ denote the worst-case subpopulation obtained by plugging the ground truth rewards $\{r_{\boldsymbol{\theta}^\star_u} : u \in [U]\}$ in (15). Accordingly, we denote the suboptimality gaps as

$$\begin{aligned} \Delta_u &\triangleq \Big| H\big(\nu^\star_{u^\star}(\cdot \mid \mathbf{X})\big) - H\big(\nu^\star_u(\cdot \mid \mathbf{X})\big)\Big|, \ \forall u \neq u^\star, \\ \Delta_{\min} &\triangleq \min_{u\in[U]\setminus\{u^\star\}} \Delta_u, \end{aligned} \tag{17}$$

where Δ_u captures the gap in the conditional entropy of the Gibbs distributions corresponding to the worst-case subpopulation $u^\star$ and any other subpopulation $u \neq u^\star$. The minimum gap $\Delta_{\min}$ captures the hardness of identifying the worst-case subpopulation under the ground truth from the other subpopulations. Furthermore, for each $u \in [U]$ we define dataset-specific parameters

$$\begin{aligned} \psi_u(\Sigma, \beta, \delta) &\triangleq \frac{1}{\beta^2}(C_\delta + B^2_{\max}) \\ &\times \left(\max_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{y}\sim\nu^\star_u(\cdot \mid \mathbf{x})}\Big[\|\phi(\mathbf{x},\mathbf{y})\|_{\left(\Sigma+\lambda\mathbb{I}\right)^{-1}}\Big]\right)^2. \end{aligned}$$

Based on these definitions, we provide the following guarantee on the sample complexity of the SharedRep-RLHF algorithm.

Theorem 2 (Sample Complexity)**.** *Let us set $\lambda = \frac{1}{N}$. Under Assumption 1,*

$$\begin{aligned} N^{\mathrm{SR}} \ \triangleq \ \max\Bigg\{ & N_{\mathrm{MaxMin}}, \\ & O\left(\frac{C_\delta}{\varepsilon^2}\big\|\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi^\star(\cdot \mid \mathbf{x})}\big[\phi(\mathbf{x},\mathbf{y})\big]\big\|^2_{\left(\Sigma+\lambda\mathbb{I}\right)^{-1}}\right)\Bigg\} \end{aligned} \tag{18}$$

samples are sufficient to ensure that $\mathbb{P}\big(J_{\mathrm{MaxMin}}(\pi^\star) - J_{\mathrm{MaxMin}}(\widetilde{\pi}^{\mathrm{SR}}) \leq \varepsilon\big) \ > \ 1-\delta$, where N_{MaxMin} is defined as follows.

1. *(Large-gap regime.) If $\Delta_{\min} > \frac{2}{\mathrm{e}}\big(\log|\mathcal{Y}|+2\big)$, we define*

$$N_{\mathrm{MaxMin}} \triangleq \max_{u\in[U]}\left\{O\left(\psi_u(\Sigma,\beta,\delta)\Big(\frac{\log|\mathcal{Y}|+2}{\Delta_{\min}}\Big)^4\right)\right\}. \tag{19}$$

2. *(Small-gap regime.) If $\Delta_{\min} \leq \frac{2}{\mathrm{e}}\big(\log|\mathcal{Y}|+2\big)$, we define*

$$\begin{aligned} N_{\mathrm{MaxMin}} \triangleq \max_{u\in[U]}\Bigg\{ & O\Bigg(\psi_u(\Sigma,\beta,\delta) \\ & \exp\Big(-4W_{-1}\Big(-\frac{\Delta_{\min}}{2(\log|\mathcal{Y}|+2)}\Big)\Big)\Bigg)\Bigg\}, \end{aligned} \tag{20}$$

where W_{-1} denotes the non-principal real branch of the Lambert-W function.

From Theorem 2, we observe that the sample complexity has two components: (1) a component that scales as $O(1/\varepsilon^2)$, and is attributed to the number of samples required for the convergence of the policy induced by parameter estimates to the ground truth $\pi^\star$, and, (2) a second component, N_{MaxMin}, which is a price we pay for the MaxMin objective, and quantifies the number of samples required to ensure that $\widehat{u} = u^\star$. Note that in the large-gap regime, the additional price scales as $O\big(1/\Delta^4_{\min}\big)$, and it may be a small price for large values of $\Delta_{\min}$. On the other hand, since $W_{-1}(-x) \approx \log x$ for small $x \in \mathbb{R}_+$, we notice that there is an $O\big(1/\Delta^4_{\min}\big)$ scaling in the small-gap regime, which may be substantial. While the first component in (18) has been reported in the literature (see, e.g., (Xiong et al. 2024)) as the sample complexity of uniform-reward RLHF, the second component is attributed to the MaxMin objective, and is a novel observation, which, to the best of our knowledge, is the first sample complexity guarantee of the MaxMin-RLHF objective.

5 Experiments

In this section, we evaluate the performance of the SharedRep-RLHF algorithm on various language tasks, comparing it against MaxMin-RLHF. Our empirical study is guided by two central questions: (1) Does SharedRep-RLHF improve group fairness over MaxMin-RLHF for tasks with low minority representation, as measured by the *average minority score*? (2) Does SharedRep-RLHF achieve a higher *win rate* on these tasks under the same conditions? We summarize the key empirical findings in this section, while deferring experimental setup details, additional results, and ablation studies to the extended version.

Tasks. We present two distinct language tasks: *controlled sentiment analysis* and *mathematical reasoning*. Additional experimental details and results for *single-turn dialogue* are deferred to the extended version.

Controlled sentiment analysis. We use the IMDb dataset (Maas et al. 2011) following Chakraborty et al., and adopt the prompt construction methodology proposed by Rafailov et al.. To simulate group-specific preferences, the dataset is randomly partitioned into *majority* and *minority* sub-populations, where the majority prefers concise reviews, whereas the minority prefers a mixture of concise (30%) and positive (70%) reviews.

Mathematical reasoning. We use the GSM8K dataset (Cobbe et al. 2021), which contains high-quality grade school math word problems requiring multi-step arithmetic reasoning. We construct prompts by appending each question to a fixed two-shot prefix designed to elicit chain-of-thought (CoT) reasoning. We define two sub-populations: the *majority*, which values brevity (80%) and correctness (20%), and the *minority*, which values correctness (20%) and *socraticity*–the extent to which the dialogue proceeds by asking guiding questions–(80%).

Min. Proportion	IMDb – Gold Score: 0.81 ± 0.001				GSM8K – Gold Score: 0.41 ± 0.003			
	Mean Minority Score		Min. Win Rate (%)		Mean Minority Score		Min. Win Rate (%)	
	MaxMin	SharedRep	MaxMin	SharedRep	MaxMin	SharedRep	MaxMin	SharedRep
0.01	0.59 ± 0.001	$\mathbf{0.69} \pm 0.009$	21.38	**25.26**	0.24 ± 0.003	$\mathbf{0.31} \pm 0.002$	19.92	**29.72**
0.05	0.63 ± 0.001	$\mathbf{0.69} \pm 0.015$	25.21	**26.01**	0.14 ± 0.002	$\mathbf{0.29} \pm 0.002$	4.93	**25.32**
0.10	0.65 ± 0.001	$\mathbf{0.71} \pm 0.013$	28.31	**31.80**	0.19 ± 0.002	$\mathbf{0.29} \pm 0.002$	11.43	**28.90**
0.15	-	-	-	-	$\mathbf{0.32} \pm 0.003$	0.28 ± 0.002	**30.52**	24.92
0.20	0.67 ± 0.001	$\mathbf{0.72} \pm 0.012$	23.97	**31.49**	$\mathbf{0.30} \pm 0.004$	0.27 ± 0.002	**31.65**	22.76

Table 1: Comparison of mean score and win rate (%) between MaxMin- and SharedRep-RLHF across different minority proportions for both IMDb ($K = 2$), and GSM8K ($K = 16$). Preference labels are obtained by comparing the gold scores for each response pair. For IMDb we score positiveness with `lvwerra/distilbert-imdb` and take conciseness to be the response length. For GSM8K we score socraticity with a custom `openai-community/gpt2-large` reward model trained on GSM8K socratic split, measure correctness using standard GSM8K evaluation, and use response length for conciseness.

Models. For the controlled sentiment analysis task, we employ `lvwerra/gpt2-imdb` as the policy model and utilize `openai-community/gpt2` as the backbone for the reward model. For the mathematical reasoning task, the policy model is `Qwen/Qwen2.5-Math-1.5B`, paired with a reward model based on `openai-community/gpt2-large`. In both settings, we incorporate (i) a single regression head for `MaxMin-RLHF`, and (ii) a linear projection followed by two regression heads to capture the *shared* and *group-specific* components in `SharedRep-RLHF`.

Results. For both tasks, we evaluate (i) the average minority-group reward obtained by the MaxMin- and SharedRep-RLHF policies on the test sets, and (ii) the win rate against corresponding "gold-reward" skylines, which are MaxMin policies trained directly on ground-truth rewards. The win rate is computed as the proportion of prompts where the evaluated policy receives a higher reward than the reference.

Controlled sentiment analysis. Table 1 reports the mean minority scores and win rates for SharedRep-RLHF and MaxMin-RLHF across varying minority group proportions on the IMDb task, along with the gold reward skyline which assumes to know the ground truth. SharedRep-RLHF consistently outperforms MaxMin-RLHF in terms of minority group satisfaction, achieving higher average scores across all proportion levels. Notably, when the minority proportion is just 1%, SharedRep-RLHF delivers a substantial 16.5% improvement over MaxMin-RLHF in mean minority score (0.687 vs. 0.590), highlighting its effectiveness in extreme imbalance scenarios. In addition to stronger performance, SharedRep-RLHF exhibits greater robustness: as the minority proportion decreases from 20% to 1%, MaxMin-RLHF suffers a 12.33% relative drop in minority score, whereas SharedRep-RLHF degrades by only 5.53%. This indicates that SharedRep-RLHF better preserves minority utility under growing data imbalance. We also compare win rates – the proportion of preference pairs in which a model's response is preferred over the gold responses. SharedRep-RLHF outperforms MaxMin-RLHF at every proportion level, with gains ranging from 0.8% to 7.5%.

Mathematical reasoning. Table 1 shows the performance of MaxMin-RLHF and SharedRep-RLHF on GSM8K across varying minority proportions, and how these compare to MaxMin-RLHF with gold rewards. SharedRep-RLHF outperforms MaxMin-RLHF in terms of mean minority score in 3 out of 5 settings, with especially large gains at low proportions: at 5% minority, the mean score improves by over 110.2% (0.288 vs. 0.137). In terms of minority win rate, SharedRep-RLHF again dominates, improving over MaxMin-RLHF by 9.8 to 20.39 percentage points. The most significant difference is at 5%, where SharedRep-RLHF increases the win rate by approximately five-fold (25.32% vs. 4.93%). These results highlight SharedRep-RLHF's ability to generalize minority-aligned generation under severe imbalance, leveraging shared traits across groups to improve estimation fidelity.

6 Conclusions

In this paper, we revisited the problem of RLHF with diverse preferences. We identified a crucial bottleneck of the state-of-the-art (Chakraborty et al. 2024) and demonstrated that MaxMin-RLHF, purported toward achieving group-fairness, suffers significantly under data imbalances. We proposed to leverage shared traits across groups to improve reward estimation, and consequently, achieve policy alignment that is more robust in terms of fairness under significant data imbalances across groups. We provided provable guarantees that SharedRep-RLHF, our proposed algorithm, outperforms MaxMin-RLHF, and characterized the sample cost of achieving group-fairness compared to uniform RLHF. Finally, we conducted comprehensive experiments on various language tasks, which showcase the performance of SharedRep-RLHF against MaxMin-RLHF, exhibiting the advantage of learning shared traits to improve reward estimation.

Acknowledgments

This work was supported in part by the UKRI for the Project AI-R (ERC Consolidator) under Grant EP/X030806/1; and by the INFORMED-AI Hub under Grant EP/Y028732/1.

References

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

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

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

Cen, S.; Mei, J.; Goshvadi, K.; Dai, H.; Yang, T.; Yang, S.; Schuurmans, D.; Chi, Y.; and Dai, B. 2024. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*.

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

Chen, D.; Chen, Y.; Rege, A.; and Vinayak, R. K. 2024. PAL: Pluralistic Alignment Framework for Learning from Heterogeneous Preferences. *arXiv:2406.08469*.

Christian, B. 2021. *The alignment problem: How can machines learn human values?* Atlantic Books.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Conitzer, V.; Freedman, R.; Heitzig, J.; Holliday, W. H.; Jacobs, B. M.; Lambert, N.; Mosse, M.; Pacuit, E.; Russell, S.; Schoelkopf, H.; Tewolde, E.; and Zwicker, W. S. 2024. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 9346–9360. PMLR.

Das, N.; Chakraborty, S.; Pacchiano, A.; and Chowdhury, S. R. 2024. Active preference optimization for sample efficient RLHF. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Dong, Y. R.; Hu, T.; Liu, Y.; Üstün, A.; and Collier, N. 2025. When Personalization Meets Reality: A Multi-Faceted Analysis of Personalized Preference Learning. *arXiv preprint arXiv:2502.19158*.

Foster, D. J.; Mhammedi, Z.; and Rohatgi, D. 2025. Is a Good Foundation Necessary for Efficient Reinforcement Learning? The Computational Role of the Base Model in Exploration. *arXiv preprint arXiv:2503.07453*.

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized Soups: Personalized Large Language Model Alignment via Post-hoc Parameter Merging. *arXiv preprint arXiv:2310.11564*.

Kong, D.; and Yang, L. 2022. Provably feedback-efficient reinforcement learning via active reward learning. *Advances in Neural Information Processing Systems*, 35: 11063–11078.

Liu, Y.; Guo, Z.; Liang, T.; Shareghi, E.; Vulić, I.; and Collier, N. 2024. Aligning with logic: Measuring, evaluating and improving logical consistency in large language models. *arXiv preprint arXiv:2410.02205*.

Maas, A.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 142–150.

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

Pang, R. Y.; Yuan, W.; He, H.; Cho, K.; Sukhbaatar, S.; and Weston, J. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37: 116617–116637.

Poddar, S.; Wan, Y.; Ivison, H.; Gupta, A.; and Jaques, N. 2024. Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning. *arXiv:2408.10075*.

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

Ramesh, S. S.; Hu, Y.; Chaimalas, I.; Mehta, V.; Sessa, P. G.; Bou Ammar, H.; and Bogunovic, I. 2024. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37: 37100–37137.

Saha, A.; Pacchiano, A.; and Lee, J. 2023. Dueling rl: Reinforcement learning with trajectory preferences. In *International conference on artificial intelligence and statistics*, 6263–6289. PMLR.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 29971–30004. PMLR.

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

Sen, A. 2017. *Collective Choice and Social Welfare: An Expanded Edition*. Cambridge, MA and London, England: Harvard University Press. ISBN 9780674974616.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Son, S.; Bankes, W.; Yoon, S.; Ramesh, S. S.; Tang, X.; and Bogunovic, I. 2025. Robust Multi-Objective Controlled Decoding of Large Language Models. *arXiv preprint arXiv:2503.08796*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2022. Learning to summarize from human feedback. arXiv:2009.01325.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Xie, T.; Foster, D. J.; Krishnamurthy, A.; Rosset, C.; Awadallah, A.; and Rakhlin, A. 2024. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-constraint. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 54715–54754. PMLR.

Yang, J.; Hu, W.; Lee, J. D.; and Du, S. S. 2020. Impact of Representation Learning in Linear Bandits. In *International Conference on Learning Representations*.

Zhang, S.; Yu, D.; Sharma, H.; Zhong, H.; Liu, Z.; Yang, Z.; Wang, S.; Hassan, H.; and Wang, Z. 2024. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*.

Zhu, B.; Jordan, M.; and Jiao, J. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, 43037–43067. PMLR.