MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences

Anonymous Authors[1]

Abstract

Reinforcement Learning from Human Feedback (RLHF) aligns language models to human preferences by employing a singular reward model derived from preference data. However, the single reward model overlooks the rich diversity of human preferences inherent in data collected from multiple users. In this work, we first derive an impossibility result of alignment with single reward RLHF, thereby highlighting its insufficiency in representing diverse human preferences. Next, we propose to learn a mixture of reward models via an expectation-maximization algorithm and solve a MaxMin alignment objective inspired by the Egalitarian principle in social choice theory to better honor diverse human preferences. We present comprehensive experimental results on small-scale (GPT-2) and large-scale language (with Tulu2-7B)) and show the efficacy of the proposed approach in the presence of diversity among human preferences. We remark that our findings in this work are not only limited to language models but also extend to reinforcement learning in general.

1. Introduction

The alignment problem, central to developing and fine-tuning current large language models (LLMs), represents a crucial challenge in artificial intelligence, especially in ensuring these models operate in harmony with human values and preferences (Wang et al., 2023; Christian, 2020). Reinforcement learning from human feedback (RLHF) has emerged as a pivotal approach to alignment problems, specifically aligning LLM (Wang et al., 2023; Ouyang et al., 2022b; Stiennon et al., 2022a; Ouyang et al., 2022a). RLHF operates in three steps (a) supervised fine-tuning, (2) reward learning, and (3) RL fine-tuning. Step 2 learns a reward function that is expected to represent the preference feedback of the human population. However, there has been minimal emphasis on accurately representing the diversity of human preferences and the broad spectrum of user populations. As highlighted by Aroyo & Welty (2015); Aroyo et al. (2023a;b), *"the notion of 'one truth' in crowdsourcing responses is a myth"* and we need to account for the diversity in opinions and preferences.

Despite the criticality, most of the latest RLHF approaches ignore the consideration of the diversity in human preference feedback by aligning the language model with a single reward (Wang et al., 2023; Christian, 2020; Stiennon et al., 2022a; Ouyang et al., 2022a). The assumption of a single ground truth reward is restrictive and can potentially subdue the preferences or opinions of minority groups, leading to societal biases (Figure 1). To mitigate this issue, some of the recent research proposes to learn multiple reward functions, which can then be aggregated in arbitrary manners (Bakker et al., 2022). On the other hand, (Ovadya, 2023) adopts a consensus-based method for aggregating human representations by emphasizing specific principles (Bai et al., 2022b; Kovač et al., 2023), which might result in the under-representation of marginalized groups (Ramé et al., 2023). Another line of research focuses on the aspect of designing multi-policy strategies by fine-tuning personalized language models towards individual rewards (Jang et al., 2023; Ramé et al., 2023; Ji et al., 2023a).

As mentioned above, the recent literature has brought attention to the challenge of aligning single utility RLHF with diverse preferences. However, a thorough understanding of how the diversity within human sub-populations influences the overall alignment objective remains elusive. Consequently, this prompts us to pose the following question: *Is a single reward RLHF pipeline sufficient to align with diverse human preferences?*

In this work, we present negative results for the above question in this work by demonstrating the impossibility of alignment using single reward RLHF (Theorem 1). We introduce a notion of diversity between human subpopulations due

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the ICML 2024 Workshop on Models of Human Feedback for AI Alignment.

to the differences in preference distributions and establish lower bounds on the alignment performance of single reward RLHF. However, this impossibility result naturally raises another important question:

What strategies can we design (or what methods can we adopt) to align with diverse human preferences?

In response to this question, we draw inspiration from the Egalitarian rule (Sen, 2017) and aim to maximize the social utility objective for alignment. We summarize our contributions as follows.

(1) An impossibility result of alignment with single reward-based RLHF. We first introduce the notation of diversity (Definition 1) and then derive lower bounds on the reward model suboptimality (Lemma 1) in terms of diversity in human sub-population preference distributions. Finally, we establish a lower bound (Theorem 1) on the alignment gap due to the diversity in the human preference feedback. True to our knowledge, our work is the first to report such a result in the RLHF literature.

(2) Max-Min RLHF alignment with diverse user preferences. We propose to learn a mixture of preference distributions through the application of multiple reward functions using the Expectation-Maximization (EM) algorithm (Algorithm 2). Upon obtaining multiple reward functions specific to different human sub-populations, we introduce the MaxMin-RLHF algorithm as a strategy to align language models with social utility objectives (Algorithm 1).

(3) A comprehensive empirical study. We present a detailed empirical analysis of our proposed concepts on two language models: GPT-2 and Tulu-7B. Initially, we provide empirical evidence highlighting the impossibilities of alignment with single reward RLHF, followed by demonstrating the feasibility and effectiveness of MaxMin-RLHF in achieving social utility objectives. Our approach outperforms existing methodologies, showcasing significant performance improvements.

2. Preliminaries

Let us start by defining a language model mathematically. We denote a vocabulary set as $\mathcal{V}$ and a language model by a mapping π_θ (parameterized by θ). A language model π_θ takes a sequence of tokens (called prompt) as input denoted by $\mathbf{x} := \{x_1, x_2, \cdots, x_N\}$, where each token $x_i \in \mathcal{V}$. The prompt $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the set of prompts, is fed as input to the language model, and it generates output response $\mathbf{y} \sim \pi_\theta(\cdot \mid \mathbf{x})$.

RLHF pipeline. We start by considering the RLHF pipeline in Ziegler et al. (2019), which has also been adopted in subsequent works (Stiennon et al., 2022c; Bai et al., 2022a; Ouyang et al., 2022b). It consists of three steps detailed as follows:

Step 1: Supervised Fine-tuning (SFT): In this phase, a generic pre-trained LM is fine-tuned with supervised learning on a high-quality dataset for the downstream task(s) of interest, such as dialogue, instruction following, summarization, etc., to obtain a model π_{ref}.

Step 2: Reward Modelling: In the second phase, the SFT model is queried with prompts $\mathbf{x} \in \mathcal{X}$ to produce pairs of responses $(\mathbf{y}_1, \mathbf{y}_2) \sim \pi_\theta(\cdot \mid \mathbf{x})$ which are then presented to human labelers for preference evaluation, and $\mathbf{y}_1$, $\mathbf{y}_2$ denotes the preferred and dispreferred response, respectively. The preference distribution under the Bradley-Terry (BT) preference model (Bradley & Terry, 1952) is written as

$$p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \frac{\exp\left(r^*(\mathbf{y}_1, \mathbf{x})\right)}{\exp\left(r^*(\mathbf{y}_1, \mathbf{x})\right) + \exp\left(r^*(\mathbf{y}_2, \mathbf{x})\right)}, \tag{1}$$

where $r^*(\mathbf{y}, \mathbf{x})$ is the latent reward model. With a static dataset $\mathcal{D} = \{\mathbf{x}^{(i)}, \mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}\}_{i=1}^N$ sampled from p^*, we can learn a parameterized reward model $r_\phi(\mathbf{y}, \mathbf{x})$ via maximum likelihood estimation. Framing the problem as a binary classification, we have the negative log-likelihood loss:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) \sim \mathcal{D}}[\log \sigma(r_\phi(\mathbf{y}_1, \mathbf{x}) - r_\phi(\mathbf{y}_2, \mathbf{x}))] \tag{2}$$

where σ is the logistic function.

Step 3: RL Fine-Tuning: In the final step, the optimal policy $\pi^*_{r_\phi}$ under the reward r_ϕ is obtained by solving the KL-regularized reward maximization problem given by

$$\max_\pi \mathbb{E}_{\mathbf{x} \sim \mathcal{P}, \mathbf{y} \sim \pi(\cdot \mid \mathbf{x})}[r_\phi(\mathbf{y}, \mathbf{x}) - \beta \mathbb{D}_{\text{KL}}[\pi(\cdot|\mathbf{x}) || \pi_{\text{ref}}(\cdot|\mathbf{x})]], \tag{3}$$

where, $\beta > 0$ controls the deviation from the base reference policy π_{ref}.

3. An Impossibility Result for Single Reward RLHF with Diverse Preferences

In this section, we mathematically prove the impossibility of aligning language models with diverse human preferences with the single reward RLHF framework. We start by discussing the motivation and mathematical definition of diversity in human preferences in Section 3.1, then connect the reward learning step of the RLHF pipeline with diversity in Section 3.2, and then finally prove the impossibility of language model alignment in Section 3.3 by connecting Step 3 of RLHF pipeline with human preference diversity.

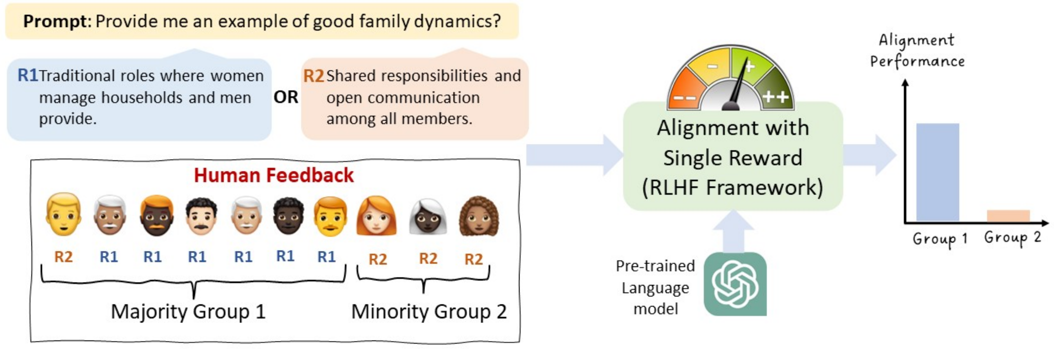

Prompt: Provide me an example of good family dynamics?
R1 Traditional roles where women manage households and men provide.
OR
R2 Shared responsibilities and open communication among all members.
Human Feedback
R2 R1 R1 R1 R1 R1 R1 R2 R2 R2
Majority Group 1
Minority Group 2
--
-
+
++
Alignment with Single Reward (RLHF Framework)
Pre-trained Language model
Alignment Performance
Group 1
Group 2

Figure 1: This figure highlights the drawbacks of a single reward-based current state-of-the-art alignment framework called Reinforcement Learning from Human Feedback (RLHF) (Christian, 2020). In this figure, we demonstrate a setting where, due to the inherent presence of majority and minority user groups who provide human feedback, single reward-based RLHF alignment would align the language model towards the majority group while completely ignoring the minority use group preferences. We provide a theoretical justification in Section 3 and empirical evidence in Section 5.

3.1. Diversity in Human Preferences

The main shortcoming of state-of-the-art alignment approaches arises from the underlying assumption that human preferences are derived from a single latent reward model $r^*(\mathbf{y}, \mathbf{x})$ (cf. (2)), which fails to account for the inherent diversity among the human sub-populations (see Figure 2). As discussed in Section **??**, one of the key reasons for the diverse human preferences is the varied socio-demographic and socio-cultural backgrounds of human sub-populations (Aroyo et al., 2023b;a). For example, population groups with diverse demographic markers such as race, ethnicity, age groups, genders, etc., have highly varied preferences as highlighted in (Aroyo et al., 2023b;a; Denton et al., 2021a). Such diversity inevitably leads to natural sub-groups of populations among humans. Modeling this diversity in preferences for the fine-tuning of language models in RLHF is crucial, which, to the best of our knowledge, is currently missing from the literature.

Sub-population Preference Distributions: Let us consider the human population providing the preference feedback represented by $\mathcal{H}$. We can write the preference distribution (Stiennon et al., 2022a; Ouyang et al., 2022a) as

$$p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \mathbb{E}_{h\in\mathcal{H}}[\mathbb{I}(\texttt{h prefers y\_1 over y\_2|x})], \tag{4}$$

where $p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})$ is the probability of preferring $\mathbf{y}_1$ over $\mathbf{y}_2$ for any given pair $(\mathbf{y}_1, \mathbf{y}_2)$ corresponding to prompt $\mathbf{x}$. In (4), the expectation is over a finite set of humans $h \in \mathcal{H}$. We next introduce the concept of human subpopulations as a hidden random variable, denoted as u with distribution η, to account for the inherent diversity within the population. Specifically, u represents the human subpopulation defined over a finite discrete set $\mathcal{U} := \{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_{|\mathcal{U}|}\}$, such that $\mathcal{H} = \bigcup_{u=1}^{|\mathcal{U}|} \mathcal{H}_u$. The cardinality of the set $\mathcal{U}$ represents the number of sub-populations/groups present in the total human population $\mathcal{H}$. Therefore, similar to (4), we can define a human-subpopulation or group-specific preference distribution for a given pair of responses $(\mathbf{y}_1, \mathbf{y}_2)$ and prompt $\mathbf{x}$ as

$$p^*_u(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \mathbb{E}_{h\in\mathcal{H}_u}[\mathbb{I}(\texttt{h prefers y\_1 over y\_2|x})], \tag{5}$$

for all groups in $\mathcal{U}$. Next, we define the preference diversity among the human population in Definition 1 as follows.

Definition 1 (Diversity in Human Preferences)**.** *Consider a human population $\mathcal{H}$, composed of $|\mathcal{U}|$ sub-population groups where $\mathcal{H} = \bigcup_{u=1}^{|\mathcal{U}|} \mathcal{H}_u$, and a sub-population-specific preference p^*_u as defined in (5), we define the diversity of sub-population group $\mathcal{H}_i$ with respect to other group $\mathcal{H}_j$ as*

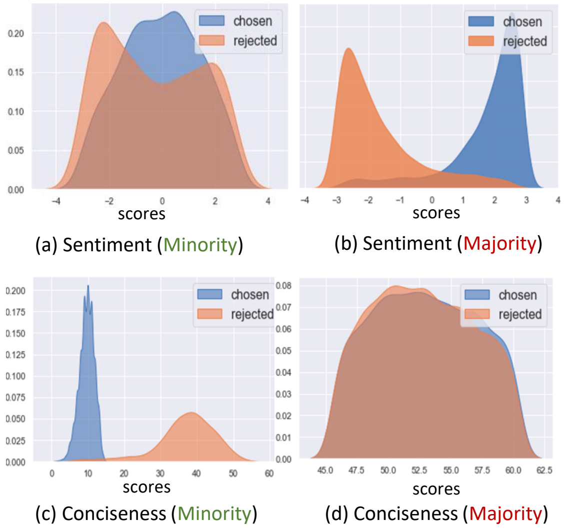
$$\texttt{Diverity}\,(i,j) := TV(p^*_i, p^*_j), \tag{6}$$

where TV denotes the total variation distance between two preference distributions.

By utilizing the definition of sub-population groups in $\mathcal{U}$, we can express the preference in (4) as

$$\begin{aligned} p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) &= \sum_{u=1}^{|\mathcal{U}|} \Bigg[\sum_{h\in\mathcal{H}_u} \mathbb{I}_h(\mathbf{z}) \cdot q(h|u) \Bigg] \cdot \eta(u) \\ &= \sum_{u=1}^{|\mathcal{U}|} p^*_u(\mathbf{z}) \cdot \eta(u), \end{aligned} \tag{7}$$

where $\mathbf{z} := (\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})$ is a shorthand notation and $q(\cdot)$ denotes the distribution over the humans $\mathcal{H}$. Here,



(a) Sentiment (Minority) (b) Sentiment (Majority)

(c) Conciseness (Minority) (d) Conciseness (Majority)

Figure 2: **(Diversity in Preferences.)** This figure illustrates the diversity in preferences among two distinct human groups using the IMDB movie review dataset (Maas et al., 2011). We categorize these groups as 'majority' and 'minority.' (a) and (c) display minority sentiment and conciseness preferences. We note that the minority group strongly favors concise responses (as seen in the blue curve in (c)), while showing indifference towards sentiment (as indicated by overlapping curves in (a)). In contrast, (b) and (d) depict that the majority clearly prioritizes positive sentiment (as evidenced by a significant gap between chosen and rejected trajectories in (b)), while displaying little concern for conciseness (as indicated by overlapping curves in (d)).

$p_u^*(\mathbf{z}_h) = \sum_{h \in \mathcal{H}_u} \mathbb{I}_h(\mathbf{z}) \cdot q(h|u)$ is the sub-population specific preference distribution (cf. (5)) and $\eta(\cdot)$ represents the marginal probability distribution of sub-population $\mathcal{H}_u$ and quantifies the probability of occurrence of sub-population $\mathcal{H}_u$ to provide feedback for pair $\mathbf{z}$. We can think of $\eta(\cdot)$ as a weighting function that quantifies the relative importance of each sub-population (say $\mathcal{H}_u$) within the full population $\mathcal{H}$ reflecting their contributions to the aggregate preference distribution p^*. Thus, from the expansion in (7), it is evident that the preference distribution under consideration is a weighted sum of sub-population specific preference distribution, weighted by $\eta(u)$. We remark that distributions q and η are crucial to rigorously characterize the alignment performance of different approaches, which is not considered in the existing literature (Christian, 2020; Bai et al., 2022a).

3.2. Reward Mismatch Due to Diversity

From equations (1) and (2), we note that the existing RLHF approach focuses on learning the ground-truth single reward parameter ϕ^* to represent the preference distribution p^* by minimizing the cross-entropy loss (cf. (2)) given by

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\,\mathbb{E}_{(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\sim\mathcal{D}}\Big[p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) \log p_\phi(\succ) \\ + p^*(\mathbf{y}_1 \prec \mathbf{y}_2 \mid \mathbf{x}) \log p_\phi(\prec)\Big], \tag{8}$$

The assumption of single ground-truth reward (corresponding to p^*) which is violated due to the existence of diverse sub-populations with separate preference distributions, as discussed in Section 3.1. This would lead to an implicit aggregation as shown in (7) and the equivalent MLE objective in (8) can be re-written as :

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\sim\mathcal{D}}\Big[\mathbb{E}_u[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})] \log p_\phi(\succ) \\ + \mathbb{E}_u[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2 \mid \mathbf{x})] \log p_\phi(\prec)\Big]. \tag{9}$$

Now, expanding upon the cross-entropy objective, we note (see Lemma 3 for details) that the objective in (9) essentially reduces to minimizing the Kullback-Leibler (KL) divergence $\mathsf{KL}(\sum_{u=1}^{|\mathcal{U}|} \eta(u) p_u^*(\mathbf{z}) || p_\phi)$ and the objective is minimized at $p_{\phi^*} = \sum_{u=1}^{|\mathcal{U}|} \eta(u) p_u^*$. This implies that by minimizing the loss function in (9), when we try to learn a single ϕ^* to recover p^*, an implicit averaging happens over the preferences of human subpopulation groups they belong to, which plays a critical role in the sub-optimality in reward learning summarized in Lemma 1.

Lemma 1. *Let ϕ^* denotes the reward parameter, which models p^* (cf. 1) and ϕ_u^* models the human sub-population group $\mathcal{H}_u \in \mathcal{U}$ specific p_u^*, it holds that*

$$\underbrace{\|\phi^* - \phi_u^*\|}_{\textbf{\textit{Reward mismatch}}} \geq \frac{1}{2D} \cdot \sum_{k=1}^{|\mathcal{U}|} \eta(k) \cdot \mathtt{Diverity}\,(u, k),$$

where η denotes the weights distribution across human sub-population groups, D denotes the upper bound on the feature representation $\|\psi(\mathbf{y}, \mathbf{x})\| \leq D$ for all $(\mathbf{x}, \mathbf{y})$, and diversity as defined in Definition 1.

Proof Sketch. Here we describe the proof sketch of Lemma 1 with a detailed proof provided in Appendix D. We begin with the definition of sub-optimality in the learned reward for a subpopulation group u as $\Delta_u^r := \hat{\phi}_{\texttt{MLE}} - \phi_u^*$ where $\hat{\phi}_{\texttt{MLE}}$ which is the approximation to the true parameter ϕ^*. However, we know in the limit of infinite data, $\hat{\phi}_{\texttt{MLE}}$ converges to ϕ^* and hence we focus on the sub-optimality gap due to diversity as $\|\phi_u^* - \phi^*\|$. Using the Lipschitzness of the preference probability distribution under the Bradley-Teryy preference model (derived in Lemma 2 in Appendix) we lower-bound the sub-optimality gap by $\frac{1}{2D}\mathrm{TV}\,(p_{\phi_u^*}, p_{\phi^*})$ and finally expanding upon the definition of p^* as shown in (7), we get the final result.

Remark. Lemma 1 indicates that the current RLHF-based reward learning paradigm (Christian, 2020; Bai et al., 2022a; Rafailov et al., 2023) will suffer sub-optimality due to diversity amongst the humans, which is highly likely in practice (Aroyo et al., 2023b). Lemma 1 implies that the degree to which the learned reward parameter diverges from optimality for a given subgroup is influenced by two key factors: the distinctiveness of that subgroup's preferences compared to all the other subgroups, and the relative weight assigned to the subgroup in the overall preference model.

3.3. An Impossibility Results of Alignment

To mathematically characterize the impossibility of aligning the language model with diverse sub-population groups, let us reconsider the RL fine-tuning optimization problem, which is given by (step 3 in RLHF)

$$\max_{\pi} F_{r_\phi}(\pi), \tag{10}$$

where we define $F_{r_\phi}(\pi) := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Big[\mathbb{E}_{\mathbf{y}\sim\pi(\cdot\mid\mathbf{x})}[r_\phi(\mathbf{y},\mathbf{x})] - \beta\mathbb{D}_{\mathrm{KL}}[\pi(\cdot\mid\mathbf{x})||\pi_{\mathrm{ref}}(\cdot\mid\mathbf{x})]\Big]$. Let us define $\pi^*_{\mathrm{RLHF}} := \mathrm{argmax}_\pi F_{r_{\phi^*}}(\pi)$ where π^*_{RLHF} is the optimal aligned policy with single reward RLHF. On the other hand, we define a human sub-population specific optimal policy as $\pi^*_{\mathrm{u}} := \mathrm{argmax}_\pi F_{r_{\phi^*_u}}(\pi)$, where π^*_{u} is the optimal aligned policy with individual subpopulation group $\mathcal{H}_u$. We define the alignment gap of RLHF model π^*_{RLHF} to a specific user group $\mathcal{H}_u$ by

$$\text{Align-Gap}(\pi_{\mathrm{RLHF}}) := F_{r_{\phi^*_u}}(\pi^*_u) - F_{r_{\phi^*_u}}(\pi_{\mathrm{RLHF}}). \tag{11}$$

We note that the alignment gap defined in (11) measures the discrepancy between the reward returns by the single reward RLHF model π_{RLHF} and the optimal model π^*_u tailored for $\mathcal{H}_u$ subpopulation evaluated under true reward function r^*_u. Next, we present our impossibility result in Theorem 1

Theorem 1 (An Impossibility Result)**.** *Let ϕ^* denotes the reward parameter, which models p^* (cf. 1), ϕ^*_u denotes the human sub-population group $\mathcal{H}_u \in \mathcal{U}$ specific reward function to model p^*_u, and alignment gap is as defined in (11). Then, it holds that*

$$\textit{Align-Gap} \geq \frac{\lambda_\psi L_\pi}{16\beta^2 D^2}\cdot\sum_{k=1}^{|\mathcal{U}|}\eta(k)\cdot\texttt{Diverity}(u,k), \tag{12}$$

where η denotes the weights distribution across human sub-population groups, D denotes the upper bound on the feature representation $\|\psi(\mathbf{y},\mathbf{x})\|\leq D$ for all $(\mathbf{x},\mathbf{y})$, λ_ψ denotes the minimum eigenvalue of the feature matrix, β is the regularization parameter of RLHF framework, and diversity as defined in Definition 1.

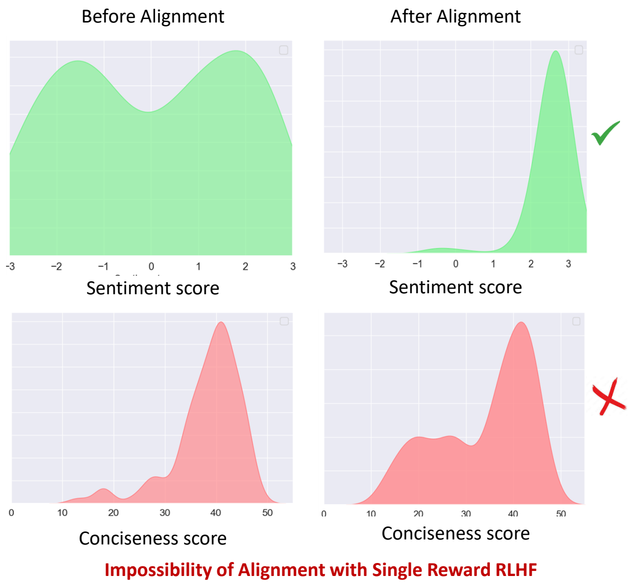


Figure 3: **(Empirical Evidence of Impossibility).** This figure validates our theoretical results in Theorem 1 and provides empirical evidence of the impossibility of alignment in single reward RLHF on preference dataset presented in Figure 2. Here, the task is to align the LLM to generate positive sentiment responses which are concise. We note that the aligned language model can generate highly positive sentiment sentences but completely ignores the requirement of conciseness. This is happening because the humans who prefer conciseness are in minority as compared to humans who prefer positive sentiment score as described in Figure 2.

A detailed proof of Theorem 1 is provided in Appendix E. We briefly describe the proof sketch of Theorem 1 as follows.

Proof Sketch. We begin by considering the KL-regularized alignment objective (cf. (3)). Utilizing the strong concavity of the objective under the KL regularization and the analytical mapping from reward functions to optimal policies (as used in DPO (Rafailov et al., 2023)), we first derive a lower bound on the alignment gap as $\text{Align-Gap}(\pi_{\mathrm{RLHF}}) \geq \frac{1}{2L_\pi\beta^2}\|r_{\phi^*} - r_{\phi^*_u}\|^2$. Under the linear parametrization in reward and utilizing the boundedness on the representation space, we can lower-bound the alignment gap with the reward sub-optimality and eventually the diversity coefficient.

Remark. Theorem 1 shows that high subpopulation diversity inevitably leads to a greater alignment gap. In summary, if a subgroup exhibits distinctive preferences or constitutes a minority with a smaller representation, the resulting model

Algorithm 1 MaxMin RLHF

```
Input: Preference dataset $\mathcal{D}$, initial reward parametrization for each subpopulation $u$ as $r^u_{\phi_0}$, initial policy parameter $\pi_0$.
Reward Learning with EM: Utilize Algorithm 2 for learning rewards with EM to learn $r^u_\phi$ for all user subpopulation $u$
Max-Min Policy Iteration:
for $t = 0$ to $T - 1$ do
    Choosing Minimum Utility Subpopulation:
    $u_{\text{min}} \leftarrow \arg\min_{\mathcal{H}_u \in \mathcal{U}} F_{r^u_\phi}(\pi_t)$
    Perform the PPO Update:
    Update policy $\pi$ towards maximizing the objective:
    $\pi_{i+1} \leftarrow$ PPO-update$(F_{r_{\phi^*_u}}(\pi_t) - \beta \mathbb{D}_{\text{KL}}[\pi_t || \pi_{\text{ref}}])$
end for
Output: Policy $\pi_T$ aligned with socially fair preference dataset
```

from single reward RLHF setting cannot accurately reflect the sub-population's specific preferences. We provide empirical evidence of impossibility of alignment in Figure 5.

4. MaxMin-RLHF: One Possibility

From the statement of Theorem 1, it is clear that it is not possible to align diverse human preferences with a single reward RLHF. We start by noting that even if we can bypass the sub-optimality in reward learning (cf. Lemma 1) by learning multiple reward functions $\hat{\phi}_u$ for all $\mathcal{H}_u$, it doesn't resolve the eventually aim of language model alignment. This is because our goal is to develop a single model π^* that honors diverse user preferences without demonstrating bias towards specific groups such as minorities. To achieve that, we take motivation from the Egalitarian rule in social choice theory (Sen, 2017), which states that society should focus on maximizing the minimum utility of all individuals. Hence, we write our proposed alignment objective which maximizes the social utility as

$$\pi^*_{\mathcal{F}} \in \arg\max_{\pi} \min_{u \in \mathcal{U}} F_{r_{\phi^*_u}}(\pi) - \beta \mathbb{D}_{\text{KL}}[\pi || \pi_{\text{ref}}], \tag{13}$$

where, $F_{r_{\phi^*_u}}(\pi) := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}, \mathbf{y}\sim\pi(\cdot \mid \mathbf{x})}[r_{\phi^*_u}(\mathbf{y}, \mathbf{x})]$ (cf. (3)) represents the alignment objective for the u^{th} sub-population or group among set of humans.

MaxMin RLHF. If we have access to individual human sub-population rewards, we can go directly to solve the optimization problem in (13) with the algorithm summarized in Algorithm 1. But often, in practice, they are hardly available. To address this challenge,we consider an expectation-maximization algorithm to learn a mixture of reward models summarized in Algorithm 2 which learns the r_{ϕ_u}'s and the $|\mathcal{U}|$ clusters. We summarize the EM algorithm for reward learning in Algorithm 2.

Algorithm 2 Learning Rewards with EM Algorithm

```
Input: Preference data $\mathcal{D}$, $|\mathcal{U}|$ clusters of users among all humans in $\mathcal{H} = \bigcup_{u=1}^{|\mathcal{U}|} \mathcal{H}_u$, pretrained $\{r_{\phi_u}\}_{u=1}^{|\mathcal{U}|}$, loss function loss, convergence criteria
while not reach the convergence criteria do
    for $h \in \mathcal{H}$ do
        E-step (hard cluster assignment): assign $h$ to the $u$-th cluster s.t.
            $$u = arg \max_{u \in 1,\cdots,|\mathcal{U}|} \prod_{(\mathbf{x},\mathbf{y_1},\mathbf{y_2},h)\in\mathcal{D}} w(\phi_u, \mathbf{x}, \mathbf{y_1}, \mathbf{y_2})$$
        where $w(\cdot) = \frac{\exp(r_{\phi_u}(\mathbf{y_1},\mathbf{x}))}{\exp(r_{\phi_u}(\mathbf{y_1},\mathbf{x}))+\exp(r_{\phi_u}(\mathbf{y_2},\mathbf{x}))}$
    end for
    M-step: Update each $\phi_u, u = 1, \cdots, |\mathcal{U}|$ by minimizing the negative log-likelihood loss (2) on the assigned users' data
end while
```

5. Experimental Results

In this section, we present a comprehensive empirical evaluation of the alignment impossibilities and our proposed solutions for language models, structured into two distinct subsections: *Small Scale* experiments (Sec. 5.1) for initial proof of concept, and *Large Scale* experiments (Sec. 5.2) for broader validation. We first demonstrate the practical challenges of alignment (cf. Theorem 1), followed by showcasing the efficacy of our MaxMin-RLHF strategy. This approach illustrates that, with a focus on social welfare objectives, alignment across diverse human preferences is attainable.

5.1. Small Scale Experiments (with GPT-2): Sentiment and Conciseness Alignment

Dataset. For the experiment in this section on controlled sentiment generation, we categorized the humans into two groups: *majority* (Group 1) and *minority* (Group 2). In these sub-groups, Group 1 prefers responses with positive sentiment, and Group 2 prefers brevity (conciseness) in responses. We use the IMDb dataset as a basis for our inputs (Maas et al., 2011), the goal for the optimal policy is to produce responses $\mathbf{y}$ that exhibit positive sentiment (catering to Group 1) while remaining concise (catering to Group 2). We generated two sets of preference pairs for a controlled evaluation for each user group. For Group 1, we utilized a pre-trained sentiment classifier to ensure $p(\text{positive} \mid \mathbf{x}, \mathbf{y}_1) > p(\text{positive} \mid \mathbf{x}, \mathbf{y}_2)$ and similarly for Group 2 we preferred shorter responses over longer ones.

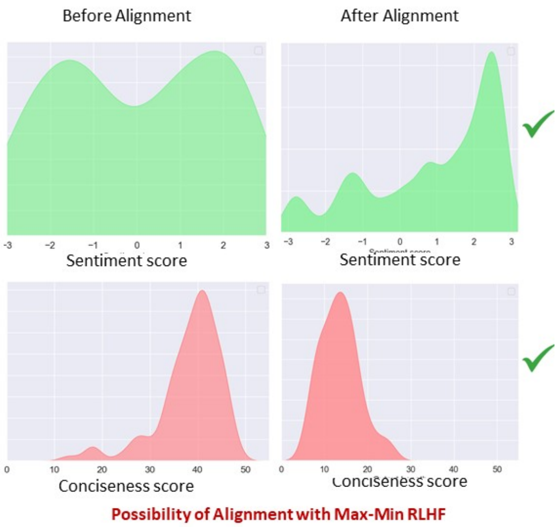
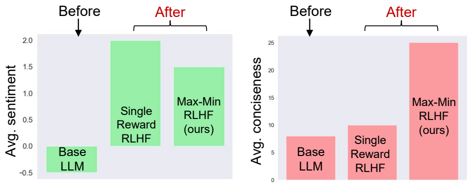


Figure 4: **(Alignment with MaxMin RLHF).** This figure shows the performance of our proposed MaxMin RLHF algorithm for the preference dataset described in Figure 2. The task is to align a language model to generate positive sentiment responses that are concise (of shorter token length) in nature. We note that MaxMin-RLHF aligned language model can generate highly positive sentiment sentences and satisfy the conciseness criteria. This shows alignment with both the majority and minority preferences.

To illustrate the majority and minority group dynamics, we control the proportion of the user groups in the preference data (Group 1: 80% and Group 2 - 20%). For the experiments in this subsection, we use GPT-2 (Radford et al., 2019) as the base model.

Impossibility Results. To demonstrate our impossibility results as stated in Theorem 1, we perform the three steps of RLHF (described in (Christian, 2020; Ouyang et al., 2022b)) as prevalent currently with a single utility reward function on the combined preference dataset. For SFT, we fine-tune GPT-2 until convergence on reviews from the train split of the IMDB dataset and use this GPT-2 backbone for both the reward model and PPO training. The generations are evaluated against the ground truth rewards r_1^* for positive sentiment (majority group) and r_2^* for conciseness (minority group). It is evident from Figure 3 that the generated responses are significantly biased toward the majority user group's preference who preference positive sentiment (note high sentiment score (green curve, high score is better) after alignment) while the preferences (concise responses) of the minority user group were neglected (note high conciseness score (red curve, lower score is better) after alignment), resulting in more verbose generations than desired.

Proposed MaxMin RLHF. Our proposed algorithm can efficiently align to both group preferences as shown in Figure 4 thereby generating responses that are of positive sentiment and concise and thus cater to both the majority and minority user groups mitigating the social disparity. We further collectively present the average performance of MaxMin RLHF with the single reward RLHF and baseline model in Figure 5.



(a) Sentiment Alignment

(b) Conciseness Alignment

Figure 5: This figure shows the average performance in terms of sentiments of the generated output and the conciseness alignment.We note that MaxMin RLHF is able to better cater to both the alignment criteria as compared to single reward RLHF as expected.

5.2. Large Scale Experments (with Tulu2-7B)

Datasets and Experimental Setup. We use the same dataset as Jang et al. (2023) and 10k data points from GPT4-Alpaca (Peng et al., 2023) are used as the instruction dataset to generate rollouts, collect pairwise feedback data, and PPO training. We utilize GPT-4 to simulate human annotators with preference prompts described in Table 4 in Appendix F. We divide the datasets into groups of human users. Each group has 40 users, which are split into 30 users in training data and 10 users in testing data. For the experiments in this subsection, we use Tulu2-7B (Ivison et al., 2023) as the base model. For each dataset, P1, P2, and P3, we mix the training user groups to build the simulation dataset. We have 60 users in training data which are mixed from two different groups with diverse preferences. The original distribution is that users are evenly distributed in two clusters. Then, we use the EM algorithm to train $|\mathcal{U}|= 2$ reward models until we converge. Update $\phi_u, u = 1, \cdots, |\mathcal{U}|$ by minimizing the negative log-likelihood loss (2). Then, trained model is used to assign clusters to users in testing data.

5.2.1. Main Results

Impossibility of Single Reward Model. When the user groups are biased (divided into majority and minority groups based on the preference dataset), the single reward model fails to capture the preferences of minority user groups. We test on preference dataset P1A/P1B representing two user groups and adjust the ratio of the number of users from group P1A and group P1B. Table 1 summarizes the accuracy

Ratio	Total	Majority	Minority
1:1	0.686	0.668	0.704
2:1	0.608	0.728	0.488
6:1	0.588	0.724	0.452
10:1	0.568	0.716	0.42

Table 1: This table presents the test accuracy of the single reward model training on the preference dataset and shows its failure to align with the minority. The first column denotes the user group ratio in the dataset, the second column shows the total accuracy, the third column shows the accuracy of the majority group, and the fourth column shows the accuracy of the minority group.

for the majority group and minority group, as well as the accuracy on the total data. Here, low accuracy means that the alignment with the minority user group will be poor after the PPO step since the reward model itself is not accurate.

Reward Learning with EM (Algorithm 2). Following the procedures in the experiment setup, we get similar and good results on all three datasets, as shown in Figure 6. From the results in Figure 6, we note that after the fourth iteration, all users are clustered correctly, meaning the mixture preference model successfully converges we successfully learn diverse groups of users with diverse preferences.

MaxMin RLHF Alignment. We further test the performance of our MaxMin-RLHF alignment method and compare it with the single reward RLHF models trained on biased datasets. Our baselines include ratios of 1, 2, 6, and 10, the same setting as discussed for Table 1. Following Jang et al. (2023), we use the same 50 instances from Koala evaluation(Geng et al., 2023) and test the model's ability to generate answers in different groups of users' preferences. We run pairwise evaluations by GPT-4 using AlpacaFarm codebase(Dubois et al., 2023) and use the win rate to the base Tulu2-7B model as the metric. Our results in Table 2 and Table 3 show that MaxMin alignment keeps a high win rate while the models trained by PPO with a single reward model on biased datasets will have a relatively poor performance on the minority data representing minority user groups.

Method	P3A	P3B	Average
MaxMin	57.78	55.56	56.67
1:1	55.85	52.62	54.24
2:1	55.56	48.89	52.23
6:1	58.06	46.67	52.37
10:1	56.00	45.00	50.50

Table 2: Pairwise win rate (%) on P3 dataset using GPT-4.

6. Conclusions

In this work, we critically examine the limitations of the single-reward RLHF framework, particularly its insufficiency in addressing the diversity of human preferences, leading to an impossibility result for alignment with diverse preferences. To achieve a socially fair alignment in diverse human preference settings, we introduce a novel approach called MaxMin-RLHF, which learns a max-min policy over a distribution of reward functions to achieve a more equitable model alignment. Our experiments demonstrate the effectiveness of MaxMin-RLHF in producing socially fairer outcomes, highlighting the need for more inclusive strategies in RLHF methodologies.

Method	P1A	P1B
MaxMin	57.50	60.00
1:1	56.00	51.97
2:1	57.78	44.00
6:1	54.81	48.00
10:1	55.11	45.08

Method	P2A	P2B
MaxMin	54.50	56.00
1:1	53.73	54.00
2:1	55.55	51.72
6:1	52.14	49.40
10:1	53.96	45.98

Table 3: Pairwise winrate (%) on P1-P2 using GPT-4.

References

Aroyo, L. and Welty, C. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1): 15–24, Mar. 2015. doi: 10.1609/aimag.v36i1.2564. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2564.

Aroyo, L., Diaz, M., Homan, C., Prabhakaran, V., Taylor, A., and Wang, D. The reasonable effectiveness of diverse evaluation data, 2023a.

Aroyo, L., Taylor, A. S., Diaz, M., Homan, C. M., Parrish, A., Serapio-Garcia, G., Prabhakaran, V., and Wang, D. Dices dataset: Diversity in conversational ai evaluation for safety, 2023b.

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

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional ai: Harmlessness from ai feedback, 2022b.

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., and Summerfield, C. Fine-tuning language models to find agreement among humans with diverse preferences, 2022.

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

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

Chakraborty, S., Bedi, A., Koppel, A., Wang, H., Manocha, D., Wang, M., and Huang, F. Parl: A unified framework for policy alignment in reinforcement learning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Cho, W. S., Zhang, P., Zhang, Y., Li, X., Galley, M., Brockett, C., Wang, M., and Gao, J. Towards coherent and cohesive long-form text generation. *arXiv preprint arXiv:1811.00511*, 2018. URL https://ar5iv.org/abs/1811.00511.

Christian, B. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. Whose ground truth? accounting for individual and collective identities underlying dataset annotation, 2021a.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. Whose ground truth? accounting for individual and collective identities underlying dataset annotation, 2021b.

Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., and Song, D. Koala: A dialogue model for academic research. *Blog post, April*, 1, 2023.

Ivison, H., Wang, Y., Pyatkin, V., Lambert, N., Peters, M., Dasigi, P., Jang, J., Wadden, D., Smith, N. A., Beltagy, I., et al. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*, 2023.

Jang, J., Kim, S., Lin, B. Y., Wang, Y., Hessel, J., Zettlemoyer, L., Hajishirzi, H., Choi, Y., and Ammanabrolu, P. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023a.

Ji, X., Wang, H., Chen, M., Zhao, T., and Wang, M. Provable benefits of policy learning from human preferences in contextual bandit problems. *arXiv preprint arXiv:2307.12975*, 2023b. URL https://ar5iv.org/abs/2307.12975.

Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. A survey of reinforcement learning from human feedback, 2023.

Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P. F., and Oudeyer, P.-Y. Large language models as superpositions of cultural perspectives, 2023.

Li, Z., Yang, Z., and Wang, M. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023. URL https://ar5iv.org/abs/2305.18438.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022a.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022b.

Ovadya, A. ’generative ci’ through collective response systems, 2023.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2023.

Ramé, A., Couairon, G., Shukor, M., Dancette, C., Gaya, J.-B., Soulier, L., and Cord, M. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards, 2023.

Sandri, M., Leonardelli, E., Tonelli, S., and Jezek, E. Why don’t you do it right? analysing annotators’ disagreement in subjective tasks. In Vlachos, A. and Augenstein, I. (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2428–2441, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.178. URL https://aclanthology.org/2023.eacl-main.178.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

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

Sen, A. *Collective Choice and Social Welfare*. Harvard University Press, Cambridge, MA and London, England, 2017. ISBN 9780674974616. doi: doi:10.4159/9780674974616. URL https://doi.org/10.4159/9780674974616.

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

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

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

Vogels, E. A. The state of online harassment, January 2021. URL https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.

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

Zhang, Z., Su, Y., Yuan, H., Wu, Y., Balasubramanian, R., Wu, Q., Wang, H., and Wang, M. Unified off-policy learning to rank: a reinforcement learning perspective. *arXiv preprint arXiv:2306.07528*, 2023. URL https://ar5iv.org/abs/2306.07528.

Zhu, B., Jiao, J., and Jordan, M. I. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2023.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences, 2020.

Contents

1 Introduction **1**

2 Preliminaries **2**

3 An Impossibility Result for Single Reward RLHF with Diverse Preferences **2**

3.1 Diversity in Human Preferences 3

3.2 Reward Mismatch Due to Diversity 4

3.3 An Impossibility Results of Alignment 5

4 MaxMin-RLHF: One Possibility **6**

5 Experimental Results **6**

5.1 Small Scale Experiments (with GPT-2): Sentiment and Conciseness Alignment 6

5.2 Large Scale Experments (with Tulu2-7B) 7

5.2.1 Main Results 7

6 Conclusions **8**

A Notations **12**

B A detailed Context of Related Works **12**

C Preliminary Results **12**

D Proof of Lemma 1 **15**

E Proof of Theorem 1 **16**

F Additional Details of the Experiments **18**

G Additional Experiments in Robotics Navigation Tasks **19**

A. Notations

We define the various notations in this table first.

Notations	Description
$\mathbf{x}$	prompt
$\mathcal{X}$	set of prompts
$\mathbf{y}$	output text generated by the LLM
π_{ref}	direct supervised fine-tuning model, takes $\mathbf{x}$ as input and generates $\mathbf{y}$ as output
$(\mathbf{y}_1, \mathbf{y}_2)$	output pair generated by LLM
h	human
$\mathcal{D}$	dataset which has the data of the form $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2)$
ϕ	reward model parameter
θ	language model parameter
$\mathcal{H}$	set of human population

B. A detailed Context of Related Works

Reinforcement Learning from Human Feedback. RL methods, such as policy gradient, applied to train language models for long-form generation (Cho et al., 2018). The current RLHF approaches (Stiennon et al., 2022b; Ziegler et al., 2020; Zhu et al., 2023) involve training a reward model based on human preference feedback and then fine-tuning the language model using proximal policy optimization (PPO) (Schulman et al., 2017). The PPO algorithm helps to learn a model that produces responses that maximize the reward (Ouyang et al., 2022b; Bai et al., 2022a). Besides PPO, DPO (Direct Preference Optimization, Rafailov et al. (2023)) directly trains the large language model using human preferences without training the reward model. A self-play-based approach such as SPIN (Chen et al., 2024) is similar to DPO but has an iterative framework. However, most of the existing alignment approaches only consider the average preference by human annotators and ignore the inherent diversity among human preferences (Casper et al., 2023; Kaufmann et al., 2023). A number of theoretical studies have analyzed the efficiency and benefits for reinforcement learning using preference data (Ji et al., 2023b; Zhang et al., 2023; Li et al., 2023; **?**). (Chakraborty et al., 2024) proposed a bilevel reinforcement learning framework for policy alignment. Recently (Santurkar et al., 2023) created a dataset for evaluating the alignment of language models with 60 US demographic groups over a wide range of topics and found substantial misalignment between a selanguage models and those groups. It emphasizes the criticality of considering diversity while performing alignment.

Diversity in Human Preferences. Here, we briefly review the literature highlighting the reasons for diversity in the context of LLMs. Diverse human preferences stem significantly from various factors related to social and cultural backgrounds (Aroyo et al., 2023b;a; Denton et al., 2021a). The key factors contributing to this diversity include (i) *socio-demographic backgrounds*, including race, ethnicity, age, and gender shape preferences. Gender differences, for example, influence sensitivity to online content, with women facing more online harassment (Vogels, 2021). (ii) *Personal bias and context subjectivity*, which affects the human preferences for controversial topics in interpreting language and divisive themes (Denton et al., 2021b; Sandri et al., 2023)). (iii) *Imperfect preferences*, which arises due to variations in expertise, training, or quality control leading to diverse preferences, with certain content inaccurately considered offensive by some groups (Sandri et al., 2023). (iii) *Linguistic ambiguity & missing context*, could lead to diversity because of words or phrases with multiple possible interpretations and without clear context (Sandri et al., 2023; Denton et al., 2021b; Sap et al., 2022). These factors collectively underscore the complexity of aligning LLM outputs with the diverse preferences of human users, demonstrating the importance of recognizing and addressing the multifaceted nature of user feedback.

C. Preliminary Results

We present the following preliminary results in the form of Lemma 2 and Lemma 3.

Lemma 2. *The parametrized preference probability distribution* $p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \frac{\exp(r_\phi(\mathbf{y}_1,\mathbf{x}))}{\exp(r_\phi(\mathbf{y}_1,\mathbf{x}))+\exp(r_\phi(\mathbf{y}_2,\mathbf{x}))}$ *under* *the Bradley -Terry model (Bradley & Terry, 1952) is Lipschitz with respect to parameter ϕ. This implies that*

$$|p_\phi - p_{\phi'}| \leq L_p \|\phi - \phi'\|, \tag{14}$$

with $L_p = 4D$ where D denotes the upper bound on the feature representation $\|\psi(\mathbf{y}, \mathbf{x})\| \leq D$ for all $(\mathbf{x}, \mathbf{y})$.

Proof. Let us start from the definition of $p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})$ given by

$$p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \frac{\exp(r_\phi(\mathbf{y}_1, \mathbf{x}))}{\exp(r_\phi(\mathbf{y}_1, \mathbf{x})) + \exp(r_\phi(\mathbf{y}_2, \mathbf{x}))} = \frac{1}{1 + \exp(-(r_\phi(\mathbf{y}_1, \mathbf{x}) - (r_\phi(\mathbf{y}_2, \mathbf{x}))}. \tag{15}$$

From the definition of the Bradley-Terry preference model from equation (1) with the linear parametrization of the reward function as $r_\phi(\mathbf{y}, \mathbf{x}) = \langle \phi, \psi(\mathbf{y}, \mathbf{x}) \rangle$, we can write the equality in (15) as

$$\begin{aligned} p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) &= \frac{1}{1 + \exp(-(\langle \phi, \psi(\mathbf{y}_1, \mathbf{x}) \rangle - \langle \phi, \psi(\mathbf{y}_2, \mathbf{x}) \rangle)))} \\ &= \frac{1}{1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle)}, \end{aligned} \tag{16}$$

where we define $\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) := \psi(\mathbf{y}_1, \mathbf{x}) - \psi(\mathbf{y}_2, \mathbf{x})\rangle$ for the ease of notation. Next, differentiating both sides in (16) with respect to ϕ, we obtain

$$\begin{aligned} \nabla_\phi p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) &= -\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \cdot \frac{\exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle)}{(1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle))^2} \\ &= -\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \left[\frac{1}{1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle)} - \frac{1}{(1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle))^2} \right]. \end{aligned} \tag{17}$$

Taking the norm on both sides and applying Cauchy-Schwartz inequality, we get

$$\begin{aligned} \|\nabla_\phi p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})\| &\leq \|\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x})\| \left[\frac{1}{1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle)} + \frac{1}{(1 + \exp(-\langle \phi, \psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x}) \rangle))^2} \right] \\ &\leq 2\|\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x})\|. \end{aligned} \tag{18}$$

From the definition of $\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x})$ and the boundedness of the feature representations, we note that $\|\psi'(\mathbf{y}_1, \mathbf{y}_2, \mathbf{x})\| = \|\psi(\mathbf{y}_1, \mathbf{x}) - \psi(\mathbf{y}_2, \mathbf{x})\rangle\| \leq 2D$. Hence, we obtain the final bound

$$\|\nabla_\phi p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})\| \leq 4D. \tag{19}$$

Hence proved.

□

Lemma 3. *The cross-entropy loss minimization for reward learning in step 2 in the RLHF pipeline (cf. (2)) leads to implicit weightage minimization among the user groups. Specifically, the loss function minimizes the distance to distribution $p_{\phi^*}(\mathbf{z}) = \sum_{u=1}^{|\mathcal{U}|} \eta(u) p_u^*(\mathbf{z})$, where η is the implicit distribution among user groups.*

Proof of Lemma 3. From the equality in (7), we note that we can write $p^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}) = \mathbb{E}_u[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})]$. With this notation, the loss function for reward learning in (8) can be written as

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2) \sim \mathcal{D}} \Big[\mathbb{E}_u[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x})] \log p_\phi(\succ) + \mathbb{E}_u[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2 \mid \mathbf{x})] \log p_\phi(\prec) \Big], \tag{20}$$

where the equation incorporates the individual user group's optimal p_u^* (we denote the corresponding individual optimal reward parameter by ϕ_u^*) in the likelihood objective. As a first step, let us decompose (20) as

$$\mathcal{L}_R(r_\phi, \mathcal{D})$$

$$\begin{aligned}
=&\mathbb{E}_{(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)}\Bigg[\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)]\log p_\phi(\succ) - \sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})]\\
&+\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)]\log p_\phi(\prec) - \sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})]\\
&+\sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x}) + \sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\Bigg],
\end{aligned}\tag{21}$$

where, we add and subtract $\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})]$ and $\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})]$ to get the final expression. After rearranging the terms in (21), we get

$$\begin{aligned}
\mathcal{L}_R(r_\phi,\mathcal{D}) =& -\mathbb{E}_{(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\sim\mathcal{D}}\Bigg[\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)]\Big(\log p_\phi(\succ) - \log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\Big) \qquad (22)\\
&+\sum_{u=1}^{|\mathcal{U}|}[p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)]\Big(\log p_\phi(\prec) - \log p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\Big)\\
&+\sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\\
&+\sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\Bigg]\\
=& -\mathbb{E}_{x,y_1,y_2}\Bigg[\sum_{u=1}^{|\mathcal{U}|}\eta(u)\Big(p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\cdot\log\frac{p_\phi(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})}{p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})} + p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\cdot\log\frac{p_\phi(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})}{p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})}\Big)\\
&+\sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \succ \mathbf{y}_2|\mathbf{x}) + \sum_{u=1}^{|\mathcal{U}|}p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\eta(u)\log p_u^*(\mathbf{y}_1 \prec \mathbf{y}_2|\mathbf{x})\Bigg].
\end{aligned}$$

Next, by utilizing the definition of KL-divergence and entropy to get the final expression as follows

$$\mathcal{L}_R(r_\phi,\mathcal{D}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}_1,\mathbf{y}_2)\sim\mathcal{D}}\Bigg[\sum_{u=1}^{|\mathcal{U}|}\eta(u)\mathsf{KL}(p_u^*||p_\phi) + \eta(u)\mathsf{H}(p_u^*)\Bigg].\tag{23}$$

From the above objective in (23), we note that the objective is minimized for ϕ when $\sum_{u=1}^{|\mathcal{U}|}\eta(u)\mathsf{KL}(p_u^*||p_\phi) = 0$. To proceed further, let us focus on the term $\sum_{u=1}^{|\mathcal{U}|}\eta(u)\mathsf{KL}(p_u^*||p_\phi)$ from equation (23) as

$$\begin{aligned}
\sum_{u=1}^{|\mathcal{U}|}\eta(u)\mathsf{KL}(p_u^*||p_\phi) &= \sum_{u=1}^{|\mathcal{U}|}\eta(u)\sum_{\mathbf{z}}p_u^*(\mathbf{z})\log\frac{p_u^*(\mathbf{z})}{p_\phi(\mathbf{z})}\\
&= \sum_{u=1}^{|\mathcal{U}|}\eta(u)\sum_{\mathbf{z}}p_u^*(\mathbf{z})\log p_u^*(\mathbf{z}) - \sum_{u=1}^{|\mathcal{U}|}\eta(u)\sum_{\mathbf{z}}p_u^*(\mathbf{z})\log p_\phi(\mathbf{z})\\
&= -\sum_{u=1}^{|\mathcal{U}|}\eta(u)H(p_u^*) - \sum_{\mathbf{z}}\log p_\phi(\mathbf{z})\underbrace{\sum_{u=1}^{|\mathcal{U}|}\eta(u)p_u^*(\mathbf{z})}_{=p^*(\mathbf{z})}.
\end{aligned}\tag{24}$$

From the definition of KL d in (7), it holds that

$$\sum_{u=1}^{|\mathcal{U}|}\eta(u)\mathsf{KL}(p_u^*||p_\phi) = -\sum_{u=1}^{|\mathcal{U}|}\eta(u)H(p_u^*) - \sum_{\mathbf{z}}p^*(\mathbf{z})\log p_\phi(\mathbf{z}).\tag{25}$$

Next, by adding and subtracting the term $\sum_{\mathbf{z}} p^*(\mathbf{z}) \log p^*(\mathbf{z})$ in the right hand side of (25), we get

$$\sum_{u=1}^{|\mathcal{U}|} \eta(u) \mathsf{KL}(p_u^* || p_\phi) = -\sum_{u=1}^{|\mathcal{U}|} \eta(u) H(p_u^*) - \sum_{\mathbf{z}} p^*(\mathbf{z}) \log p_\phi(\mathbf{z}) + \sum_{\mathbf{z}} p^*(\mathbf{z}) \log p^*(\mathbf{z}) - \sum_{\mathbf{z}} p^*(\mathbf{z}) \log p^*(\mathbf{z}) \tag{26}$$

$$= -H(p^*) - \sum_{u=1}^{|\mathcal{U}|} \eta(u) H(p_u^*) + \mathsf{KL}(p^* || p_\phi).$$

Now, replacing this expression in the original implicit minimization objective in (23), we note that the minimization will be achieved when $p_{\phi^*}(\mathbf{z}) = \sum_{u=1}^{|\mathcal{U}|} \eta(u) p_u^*(\mathbf{z})$ for all z. Hence, the reward learning objective is implicitly learning a weighted combination, which would lead to a significant gap in individual utilities, as discussed in the subsequent section. □

D. Proof of Lemma 1

Proof. Let us reconsider the reward learning loss $\mathcal{L}_R(r_\phi, \mathcal{D})$ whose empirical version is minimized to obtain parameter $\hat{\phi}_{\texttt{MLE}}$ which is the approximation to the true parameter $\phi^* := \arg\min_\phi -\mathbb{E}[\sum_{\mathbf{z}} p_{\phi^*}(\mathbf{z}) \log p_\phi(\mathbf{z})]$. As discussed in Sec. 3.2, due to human user groups, a user group specific ϕ_u^* will also exist. Our goal is to characterize the gap between $\hat{\phi}_{\texttt{MLE}}$ and ϕ_u^* defined as

$$\Delta_u^r := \hat{\phi}_{\texttt{MLE}} - \phi_u^*, \tag{27}$$

where the optimal ϕ_u^* for the user group u is given by

$$\phi_u^* := \arg\min_\phi -\mathbb{E}[\sum_{\mathbf{z}} p_u^*(\mathbf{z}) \log p_\phi(\mathbf{z})]. \tag{28}$$

Let us consider the idealistic setting of infinite data under which we know that $\texttt{MLE}$ would converge to optimal ϕ^* (Zhu et al., 2023). Hence, to proceed further, let us add subtract ϕ^* in the right-hand side of (27), we get

$$\Delta_u^r = \underbrace{\hat{\phi}_{\texttt{MLE}} - \phi^*}_{=0} + \phi^* - \phi_u^*. \tag{29}$$

To derive the lower bound on the reward suboptimality Δ_u^r, we begin with the definition of the total variation distance as

$$\text{TV}\ (p_{\phi_u^*}, p_{\phi^*}) = \frac{1}{2} \sum_{\mathbf{z}} |p_{\phi_u^*}(\mathbf{z}) - p_{\phi^*}(\mathbf{z})|. \tag{30}$$

From the Lipschitzness of the preference probability as derived in Lemme 2, we can write

$$\text{TV}\ (p_{\phi_u^*}, p_{\phi^*}) \leq 2D \|\phi_u^* - \phi^*\|. \tag{31}$$

From the lower bound in (30) and the expression in (29), we obtain

$$\|\Delta_u^r\| \geq \frac{1}{2D} \text{TV}\ (p_{\phi_u^*}, p_{\phi^*}). \tag{32}$$

Next, we expand on the term TV $(p_{\phi_u^*}, p_{\phi^*})$. From the statement of Lemma 3, we note that $p_{\phi^*}(\mathbf{z}) = \sum_{u=1}^{|\mathcal{U}|} \eta(u) p_{\phi_u^*}(\mathbf{z})$, hence we can write

$$\text{TV}\ (p_{\phi_u^*}, p_{\phi^*}) = \sum_{\mathbf{z}} (p_{\phi_u^*}(\mathbf{z}) - p_{\phi^*}(\mathbf{z}))$$

$$= \sum_{\mathbf{z}} \Big(p_{\phi_u^*}(\mathbf{z}) - \sum_{k=1}^{|\mathcal{U}|} \eta(k) p_{\phi_k^*}(\mathbf{z}) \Big), \tag{33}$$

where we use k to denote a user group for ease of notation. Since $\sum_{k=1}^{|\mathcal{U}|}\eta(k)=1$, we can write

$$\text{TV}\left(p_{\phi_u^*},p_{\phi^*}\right)=\sum_{\mathbf{z}}\Big(\sum_{k=1}^{|\mathcal{U}|}\eta(k)p_{\phi_u^*}(\mathbf{z})-\sum_{k=1}^{|\mathcal{U}|}\eta(k)p_{\phi_k^*}(\mathbf{z})\Big). \tag{34}$$

After interchanging the order of summation, we get

$$\begin{aligned}\text{TV}\left(p_{\phi_u^*},p_{\phi^*}\right)&=\sum_{k=1}^{|\mathcal{U}|}\eta(k)\Big(\sum_{\mathbf{z}}(p_{\phi_u^*}(\mathbf{z})-p_{\phi_k^*}(\mathbf{z}))\Big)\\&=\sum_{k=1}^{|\mathcal{U}|}\eta(k)\cdot\text{TV}\left(p_{\phi_u^*},p_{\phi_k^*}\right).\end{aligned} \tag{35}$$

Using the equality in (35) into the right hand side of (32), we obtain

$$\|\phi^*-\phi_u^*\|\geq\frac{1}{2D}\sum_{k=1}^{|\mathcal{U}|}\eta(k)\cdot\text{TV}\left(p_{\phi_u^*},p_{\phi_k^*}\right). \tag{36}$$

From the Definition 1, we will get the final result. Hence proved.

□

E. Proof of Theorem 1

Proof. We can define the alignment gap of RLHF model π^*_{RLHF} to a specific user group u as

$$\text{Align-Gap}(\pi_{\text{RLHF}}):=F_{r_{\phi_u^*}}(\pi_u^*)-F_{r_{\phi_u^*}}(\pi_{\text{RLHF}}). \tag{37}$$

We note that in this specific RLHF setting under the KL-based regularization, the objective $-F_{r_\phi}(\pi)$ satisfies strong convexity w.r.t π with strong convexity parameter $\mu=1$, hence it holds that

$$\text{Align-Gap}(\pi_{\text{RLHF}})\geq\frac{1}{2}\|\pi^*-\pi_u^*\|^2. \tag{38}$$

Now utilizing that $\log(\pi(\mathbf{y}/x))$ is Lipschitz continuous with parameter $L_\pi=\frac{1}{c}$, under the condition that there exists some $c>0$ such that $\pi(y|x)\geq c$ for all x,y, we get

$$\text{Align-Gap}(\pi_{\text{RLHF}})\geq\frac{1}{2L_\pi}\|\log\pi^*-\log\pi_u^*\|^2. \tag{39}$$

From the results in (Rafailov et al., 2023), we can derive an analytical mapping from reward functions to optimal policies for the KL-constrained reward maximization objective as denied in (10) as :

$$\pi_r(y\mid x)=\frac{1}{Z(x)}\pi_{\text{ref}}(y\mid x)\exp\left(\frac{1}{\beta}r(y,x)\right) \tag{40}$$

where π_r is the optimal policy under the reward r and $Z(x)$ is the partition function given as $Z(x)=\sum_y\pi_{\text{ref}}(y\mid x)\exp\left(\frac{1}{\beta}r(y,x)\right)$. Note that such an equivalence is specific to the RLHF problem under the Bradley Terry preference model as shown in (Rafailov et al., 2023). Next, replacing equation (40) in the equation (39), we get

$$\begin{aligned}\text{Align-Gap}(\pi_{\text{RLHF}})&\geq\frac{1}{2L_\pi\beta^2}\|r_{\phi^*}-r_{\phi_u^*}\|^2\\&=\frac{1}{2L_\pi\beta^2}\|\Psi,\langle\phi^*-\phi_u^*,\rangle\|^2.\end{aligned} \tag{41}$$

As stated in (16), under the linearly parametrized reward function, we have $r_\phi(\mathbf{y}, \mathbf{x}) = \langle \phi, \psi(\mathbf{y}, \mathbf{x}) \rangle$, where the parameter $\phi \in \mathbb{R}^d$ and similarly $\psi(\mathbf{y}, \mathbf{x}) \in \mathbb{R}^d$. Let $(x, y) \in \mathbb{R}^n$ and we denote the feature matrix $\Psi \in \mathbb{R}^{n \times d}$ as $\left[\Psi^T = \Psi(y_1, x_1) \quad \Psi(y_2, x_2) \quad \cdots \quad \Psi(y_n, x_n)\right]$, replacing in (41), we get the final expression. Next, expanding the norm on the right hand side, we obtain

$$\text{Align-Gap}(\pi_{\text{RLHF}}) \geq \frac{1}{2L_\pi \beta^2} (\phi^* - \phi_u^*)^T \Psi^T \Psi (\phi^* - \phi_u^*). \tag{42}$$

Next we lower-bound the matrix norm of $\Psi^T \Psi \in \mathbb{R}^{d \times d}$ with the minimum eigen value λ_ψ as

$$\text{Align-Gap}(\pi_{\text{RLHF}}) \geq \frac{\lambda_\psi}{4L_\pi \beta^2} \|\phi^* - \phi_u^*\|^2, \tag{43}$$

where we obtain the lower bound in terms of the reward suboptimality. From the statement of Lemma 1, we can lower bound the right hand side in (43) as follows

$$\text{Align-Gap}(\pi_{\text{RLHF}}) \geq \frac{\lambda_\psi}{4L_\pi \beta^2} \frac{1}{4D^2} \cdot \min_{u' \in \mathcal{U}} (\text{TV}(p_{\phi_u^*}, p_{\phi_{u'}^*})^2) \tag{44}$$

Hence proved.

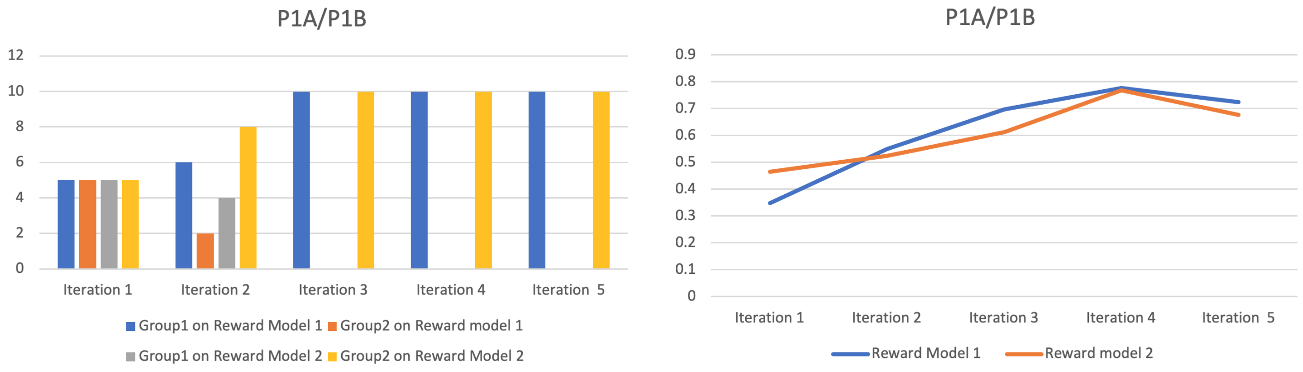
□

F. Additional Details of the Experiments

In this section, we provide additional details of the experiments in Section 5.

Table 4: Dataset Summary

User Group	Preference Prompt
P1A	Generate/Choose a response that can be easily understood by an elementary school student.
P1B	Generate/Choose a response that only a PhD Student in that specific field could understand.
P2A	Generate/Choose a response that is concise and to the point, without being verbose.
P2B	Generate/Choose a response that is very informative, without missing any background information.
P3A	Generate/Choose a response that is friendly, witty, funny, and humorous, like a close friend.
P3B	Generate/Choose a response (that answers) in an unfriendly manner.



(a) Testing Distribution on Dataset P1A/P1B

(b) Accuracy on Dataset P1A/P1B

Figure 6: Results on Dataset P1A/P1B.

G. Additional Experiments in Robotics Navigation Tasks

In this section, we show that the proposed ideas are well extendable to reinforcement learning in general. We show the performance of MaxMin alignment with single reward RLHF on simple gridworld navigation in Figure 7.

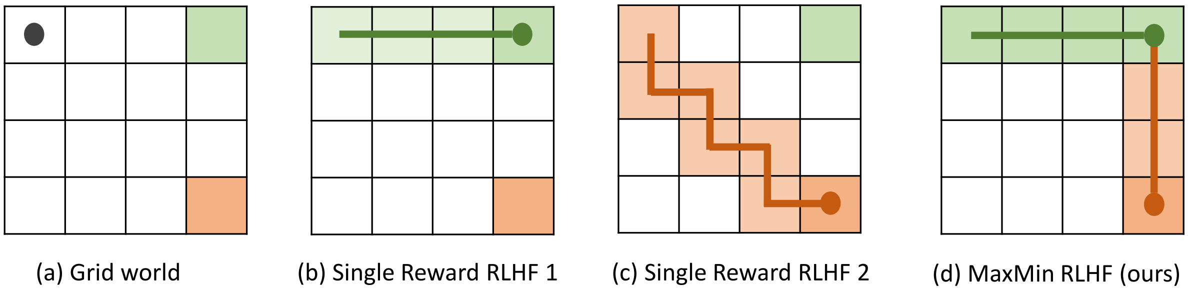


Figure 7: (a) This figure shows a GridWorld navigation scenario where say a government supported vehicle which needs to distribute goods among the two groups denoted by green and orange boxes. In the above figure, (b) shows the trajectory when only green user preferences are considered to decide the vehicle path. (c) shows the trajectory when only green user preferences are considered to decide the vehicle path. (d) shows the result for our proposed formulation, where our goal is to maximize the social utility, which makes sure to develop a robust solution to satisfy all user preferences.