Not All Voices Are Rewarded Equally: Probing and Repairing Reward Models across Human Diversity

Zihao Li, Feihao Fang, Xitong Zhang, Jiaru Zou,
Zhining Liu, Wei Xiong, Ziwei Wu, Baoyu Jing, Jingrui He

University of Illinois Urbana-Champaign
Correspondence: zihaoli5@illinois.edu, jingrui@illinois.edu

Abstract

The advancement of Large Language Models (LLMs) has made ensuring their trustworthiness increasingly critical, especially in terms of fairness across diverse human groups. While modern LLMs are aligned with user preferences through Reinforcement Learning from Human Feedback (RLHF), the reward models used for alignment are trained on preference data that may both reflect societal biases and suffer from demographic skewness, as labeler populations are often uneven due to systemic accessibility or participation gaps. In this work, we reveal that reward models can exhibit significant discrepancies across different demographic groups, posing a fundamental challenge to fair and robust alignment. Using real-world datasets, we conduct the most comprehensive study to date, auditing various state-of-the-art reward models across nine sensitive attributes, including age, gender, ethnicity, etc. Our evaluation spans both (1) the agreement level between reward models and specific user groups, and (2) the reward model's preference toward responses associated with different groups. Based on these findings, we propose the first method to mitigate group disparities in reward modeling. Code is available at `https://github.com/Violet24K/FaRM`.

1 Introduction

> "If a man knows not to which port he sails, no wind is favorable."
>
> — *Seneca, Letters to Lucilius*[1]

Before deployment in real-world applications, the vessel of Large Language Models (LLMs) is set adrift after pretraining, propelled further by the wind of Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). This phase aligns LLMs toward the destination port charted by the Reward Model's preferences. In modern LLM pipelines (Wolf et al., 2019; Dong et al., 2024; Diao et al., 2024), RLHF has become a standard step for enhancing trustworthiness, including reducing toxic outputs, refusing harmful requests, and promoting better adherence to human ethical norms (Kaufmann et al., 2023).

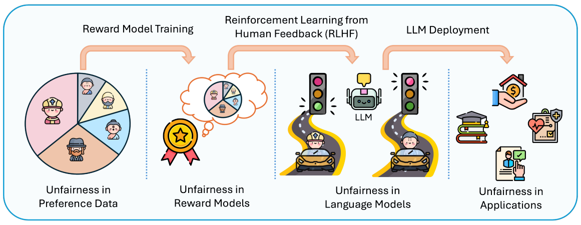

Reward Model Training
Reinforcement Learning from Human Feedback (RLHF)
LLM Deployment
LLM
Unfairness in Preference Data
Unfairness in Reward Models
Unfairness in Language Models
Unfairness in Applications

Figure 1: Pathways of Unfairness in LLM Alignment. This work focuses on **biases in reward models**, which propagate downstream into LLM applications.

In this work, we focus on one important aspect of LLM trustworthiness — fairness, which is about ensuring that models perform equitably across different demographic groups without systematically favoring or disadvantaging any particular population (Caton and Haas, 2024; Mehrabi et al., 2022). It has been demonstrated that LLMs exhibit notable unfairness across demographic groups (Li et al., 2023b), and even widely deployed models such as GPT-3.5 and GPT-4 (OpenAI, 2023) have been shown to encode substantial performance disparities across sex and race (Wang et al., 2023; Zhang et al., 2025, 2024b).

However, while modern LLMs are trained through unsupervised pretraining, supervised finetuning, and RLHF (Dubey et al., 2024; Minaee et al., 2024; Tie et al., 2025), most existing work on LLM fairness has been limited to examining the first two stages (Ghanbarzadeh et al., 2023; Yu et al., 2023; Thakur et al., 2023; Fatemi et al., 2023; Li et al., 2024) or evaluating fairness of only the final deployed LLMs (Durmus et al., 2023; Kotek et al., 2023; Myung et al., 2024). Only recently has a seminal study (Ryan et al., 2024) revealed

[1] `https://en.wikipedia.org/wiki/Seneca_the_Younger`

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 3426–3455
November 4-9, 2025 ©2025 Association for Computational Linguistics

Table 1: Comparison between this work and representative recent empirical/benchmark studies.

Reference	Venue	Sensitive Attribute Coverage				Reward Model		Bias Mitigation
		# attributes	correlation study	real data	synthetic data	# models	# model sizes	
(Ryan et al., 2024)	ACL 2024 (main)	2	✗	✓	✗	4	1 (7B)	✗
(Mire et al., 2025)	NAACL 2025 (findings)	1	✗	✗	✓	17	5 (2B-20B)	✗
(Song et al., 2025)	Preprint	1	✗	✗	✓	8	4 (7B-13B, 340B)	✗
(Ouyang et al., 2025)	ACL 2025 (main)	Focus on general language modeling, not demographic fairness				2	2 (7B-8B)	✓
Ours	-	**9**	✓	✓	✓	9	**6** (2B-27B)	✓

that RLHF introduces additional unintended geolocation bias into LLMs. Specifically, the alignment process unequally affects the model's performance across different geolocation groups, leading to unfair overall outcomes after preference tuning.

The standard RLHF pipeline (Ouyang et al., 2022) consists of three stages: preference data collection, reward modeling, and model optimization using RL from the learned reward. Consequently, the *final model behavior* is heavily influenced by the *quality of the reward model*, and, by extension, the *collected human preference data*, which may itself encode societal biases and demographic skewness. The effectiveness of RLHF fundamentally depends on the quality of reward model, trained from human preference data that may itself reflect societal biases and demographic skewness. As noted by (Ryan et al., 2024), it remains challenging to answer whose preferences are being aligned during reward model training. This difficulty arises because reward model developers retain significant control over critical but often undocumented/untraceable alignment variables, such as who provides feedback. Nevertheless, systematically understanding how reward models behave across different demographic groups is a crucial step toward identifying and mitigating potential biases.

To this end, this work presents the most comprehensive study to date on demographic biases encoded in reward models (see Table 1 for a comparison with existing studies), covering a wide range of sensitive attributes. We begin by contextualizing the notion of *fairness* in the context of RLHF reward modeling, and then proceed with a series of probing analyses. As illustrated in Figure 2, leveraging real-world interactions between users from different demographic groups and LLMs (Kirk et al., 2024; Pistilli et al., 2024), we first statistically evaluate whether different demographic groups demonstrate different preferences. We then assess how well the reward model aligns with each group and quantify disparities across sensitive attributes. After that, we explicitly evaluate the reward model's preferences on representative responses associated with each group when answering controversial questions (Li et al., 2024; Xu et al., 2024b). Finally, we propose the first reward model bias mitigation method by fine-tuning the reward model with a regularization term encouraging equal grounding across responses from different demographic groups. Our contributions toward improving fairness in reward modeling are threefold:

- We contextualize fairness in reward modeling, formalizing group-level fairness metrics.
- We conduct the most comprehensive study to date, auditing alignment disparities across demographic groups and introducing a new methodology for evaluating group-representative preferences.
- We propose the first mitigation method *FaRM* for reward model debiasing and empirically demonstrate its effectiveness.

2 Preliminary and Background

We use calligraphic letters (e.g., $\mathcal{A}$) for sets and bold capital letters for matrices (e.g., $\mathbf{A}$). For matrix indices, we use $\mathbf{A}[i, j]$ to denote the entry in the i^{th} row and the j^{th} column.

Preference Data and Reward Modeling. Modern reward models (Zhong et al., 2025; Frick et al., 2024) are trained on preference datasets $\mathcal{D} = \{x, a^w, a^l\}$, where a^w and a^l are two responses to prompt x, with a^w preferred over a^l by a human annotator. The reward models are expected to assign a higher score to the preferred response, i.e., $r(x, a^w) > r(x, a^l)$, thereby learning to approximate human preferences through pairwise comparisons. The mainstream approach (Ouyang et al., 2022) trains reward models (parametrized by θ) by maximizing the log-likelihood of Bradley-Terry (BT) model (Bradley and Terry, 1952):

$$\ell_{\mathcal{D}}(\theta) = \sum_{(x,a^w,a^l)\in\mathcal{D}} \log\Big(\sigma\big(r_\theta(x, a^w) - r_\theta(x, a^l)\big)\Big)$$

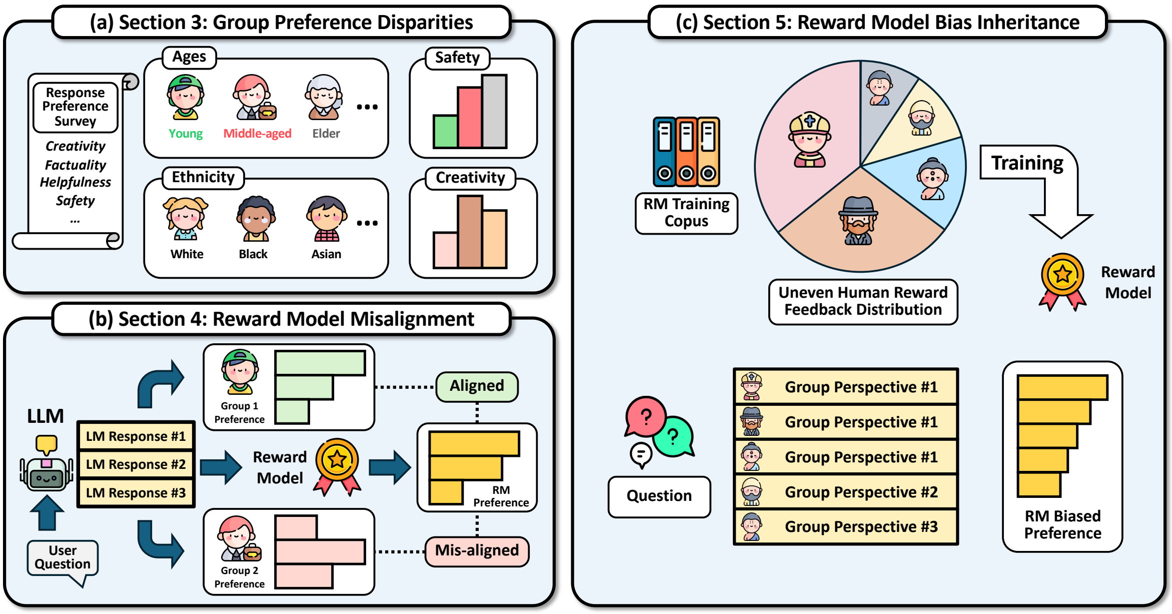


Figure 2: Overview of probing reward model bias in this study. In Section 3, we investigate how user preferences differ across demographic groups partitioned by sensitive attributes such as age and ethnicity. In Section 4, we assess how well the reward model aligns with each group by comparing the model's preferences with those of the groups. In Section 5, we directly evaluate the reward model's preferences on representative group responses, revealing inherent biases resulting from training on uneven human reward feedback.

The trained reward model r_θ then provides learning signals during subsequent RL phases.

Machine Learning Fairness. Fairness has been a long-standing goal in ML, with extensive research from traditional classifiers to modern models including language models (Liu et al., 2025, 2024b; Doan et al., 2024). Various fairness metrics have been proposed to assess disparities across demographic groups (Garg et al., 2020), including statistical parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016), and calibration within groups (Kleinberg et al., 2017). For instance, given a set of demographic groups $\mathcal{G}$, statistical parity expects that the output of the model parametrized by Θ is statistically independent of the sensitive attribute A:

$$\mathbb{E}[f_\Theta(X)|A = g_i] \approx \mathbb{E}[f_\Theta(X)|A = g_j], \forall g_i, g_j \in \mathcal{G}$$

In section 4.1, we generalize the fairness concepts to reward modeling settings.

3 How Preferences Vary Across Demographic Groups?

We begin by presenting a quantitative analysis of how user preferences vary across demographic groups. This analysis leverages the PRISM Alignment Dataset (Kirk et al., 2024) (details in Appendix C.1), which contains real-world interactions from over 1,500 participants worldwide. Each participant completed a survey reporting: (1) their demographic attributes, including age, gender, employment status, education level, marital status, English proficiency, religion, ethnicity, and geographic location (see Table 5); and (2) their stated preference across various LLM response quality dimensions, including values, creativity, fluency, factuality, diversity, safety, personalization, helpfulness, and other criteria. Preferences were elicited from each participant by scoring every dimension through a visual analog scale and explicitly ranking the dimensions (example in Table 6). Combining participants' demographic attributes with their stated preferences, we investigate whether different demographic groups exhibit distinct preference patterns. Such variation may indicate potential bias in the composition of RLHF datasets, where systemic factors, such as regional wage disparities affecting labeler participation, could lead reward models to disproportionately reflect the preferences of over-represented groups.

For a sensitive attribute (e.g., gender) with value space $\mathcal{S}$ containing all possible attribute values (e.g., male, female, non-binary, etc.), let $\mathcal{U}_s$ denote all users with sensitive attribute value $s \in \mathcal{S}$ (i.e., a demographic group), and $\mathcal{P}_s = \{p_u\}_{u \in \mathcal{U}_s}$ denote their stated preferences from range 0 to

100. Given a sensitive attribute which partitions all users into different demographic groups $\{\mathcal{U}_s\}_{s\in\mathcal{S}}$, for each demographic group $\mathcal{U}_s$, we compute the group-level average for each response quality dimension as: $\frac{1}{|\mathcal{U}_s|}\sum_{u\in\mathcal{U}_s} p_u[dim]$, where $dim \in \{\text{values}, \text{helpfulness}, \text{safety}, ...\}$ corresponds to a response quality axis. We visualize the aggregated preferences across demographic groups using radar charts in Figure 3, which illustrates group-level preferences by the sensitive attribute *ethnicity*.

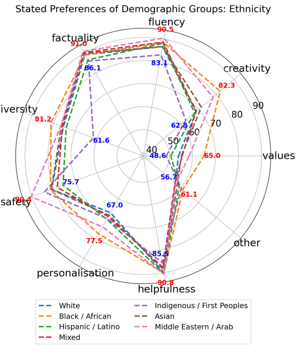


Figure 3: Stated preferences of different ethnicity groups across various LLM response quality dimensions (Kirk et al., 2024). Significance levels validated in Table 11. Full plots are in Appendix F.1.

We observe that some demographic groups exhibit significantly different preference patterns along certain dimensions. For example, in Figure 3, Black/African participants tend to assign higher scores to creativity, diversity, and personalization. This may be influenced by historical and cultural factors, including a strong tradition of creative expression in Black communities. Full figures and analysis can be found in Appendix F.1. To assess the statistical significance of preference differences across demographic groups, we perform Kruskal–Wallis tests (details in Appendix E.3) on each response quality dimension and report the resulting p-values in Table 11. To validate the robustness of the findings, rather than relying only on the stated preferences, we also performing Kruskal–Wallis tests on performance preferences (Table 12) and choice preferences (Table 13) in the PRISM dataset (Kirk et al., 2024). The results indicate that while all studied sensitive attributes exhibit some degree of preference variation, the most pronounced differences are associated with religion, geographic location, and ethnicity. On the response quality axes, the most pronounced differences are observed in values, creativity, safety, and diversity.

To identify which groups exhibit the most divergent preferences, we compute the max-min score difference for each attribute–dimension pair $(\mathcal{S}, d)$

$$\underset{s_i, s_j\in\mathcal{S}}{\operatorname{argmax}} \left| \frac{1}{|\mathcal{U}_{s_i}|}\sum_{u\in\mathcal{U}_{s_i}} p_u[d] - \frac{1}{|\mathcal{U}_{s_j}|}\sum_{u\in\mathcal{U}_{s_j}} p_u[d] \right|$$

where $\mathcal{U}_s$ denotes the set of users belonging to group $s \in \mathcal{S}$, and $p_u[d]$ is the user u's stated preference on dimension d. We visualize the resulting discrepancies in Figure 9 in the Appendix, highlighting max-min pair (s_i, s_j) with the greatest difference for each case. Darker red shades indicate more substantial divergence. We observe that certain groups, including individuals under 24 or over 55 years old, unemployed or retired individuals, followers of folk religions, Black/African participants, and users located in Africa, tend to diverge more sharply in their stated preferences.

We summarize the key takeaways from this section as follows. Based on our analysis of real user survey data, certain demographic groups exhibit notably different preference patterns across various response quality dimensions. These differences suggest that alignment based on aggregated or unbalanced preference data may disproportionately reflect the values of overrepresented groups.

4 How Reward Models Capture Group Preferences?

Given the observed differences in preferences across demographic groups, it is essential to examine how reward models capture, or fail to capture, this diversity. Specifically, we aim to further investigate the question posed by (Ryan et al., 2024): *Whose preferences are we aligning LLMs with, and, more crucially, whose preferences are overlooked?*

Although various fairness metrics have been proposed (Chouldechova and Roth, 2020), the notion of fairness in the context of preference modeling remains nascent in the existing literature. In (Ryan et al., 2024), the authors use the performance gap of the same language model before and after RLHF alignment as an implicit indicator of how well the reward model used for alignment agrees with a demographic group. In this work, we present a more direct measurement of reward model fairness.

4.1 Fairness in Reward Modeling

Let $\mathcal{X}_u$ denote a set of prompts from user u. For each prompt $x \in \mathcal{X}_u$, let $\mathcal{A}_x = \{a_1, a_2, ..., a_K\}$ be a set of responses and the user u provides a preference vector $q_u(x) \in \mathbb{R}^K$ over $\mathcal{A}_x$ where higher values indicate stronger preference. A reward model

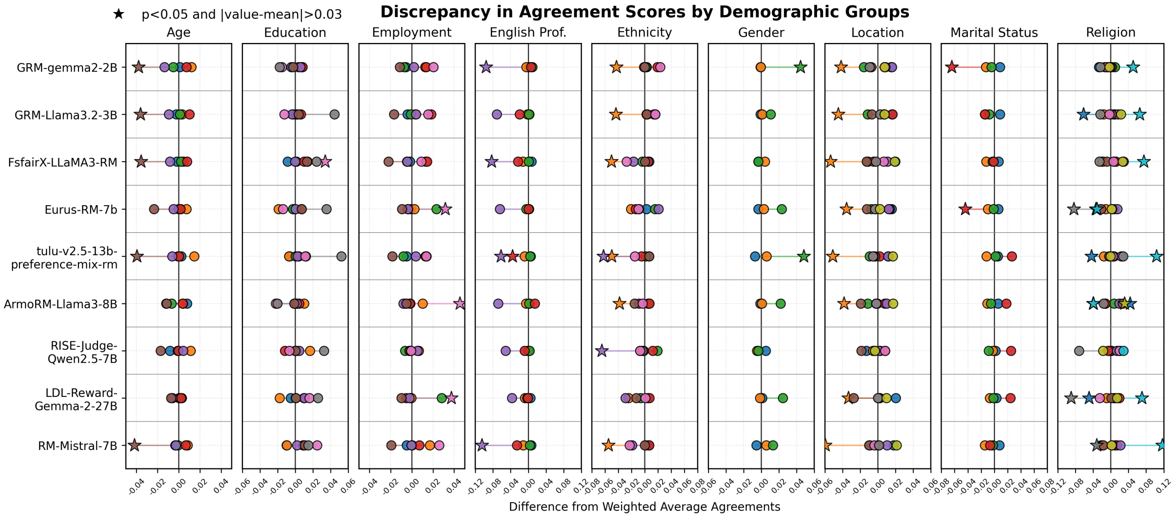


Figure 4: Different demographic groups exhibit varying levels of alignment with reward models. The values represent deviations from the average score, with stars marking significantly unfair treatments. Groups within the same sensitive attribute are indicated by the same color.

parametrized by θ assigns a score $r_\theta(x, a_i) \in \mathbb{R}$ to each response a_i. Similar as before, let $\mathcal{S}$ be the set of demographic groups for a sensitive attribute, and $\mathcal{U}_s$ denote the set of users in group $s \in \mathcal{S}$.

Reformulation to Pairwise Preference Representation. To evaluate how well a reward model captures user preferences, we reformulate each ranking into multiple pairwise comparisons, which reduces ranked preference comparison to binary classifications over response pairs. Specifically, the user u's pairwise preference of response pair (a_i, a_j) is defined as

$$y_{u,x}^{(i,j)} = \begin{cases} 1 & \text{if } q_u(x)[i] > q_u(x)[j] \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Similarly, reward model's pairwise preference is

$$\hat{y}_{\theta,x}^{(i,j)} = \begin{cases} 1 & \text{if } r_\theta(x, a_i) > r_\theta(x, a_j) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Agreement Rate. To assess how closely a reward model's preferences align with those of individual users, we measure the agreement between user-provided pairwise labels $y_{u,x}^{(i,j)}$ and the model's predicted pairwise rankings $\hat{y}_{\theta,x}^{(i,j)}$. Specifically, agreement between user u and reward model θ over response pair (a_i, a_j) is defined as:

$$\alpha_{u,\theta,x}^{(i,j)} = \begin{cases} 1 & \text{if } y_{u,x}^{(i,j)} = \hat{y}_{\theta,x}^{(i,j)} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Then, the group-level agreement rate for group $s \in \mathcal{S}$ is

$$\alpha(s) = \frac{1}{Z_s} \sum_{u \in \mathcal{U}_s} \sum_{x \in \mathcal{X}_u} \sum_{i<j \le K_x} \alpha_{u,\theta,x}^{(i,j)} \tag{4}$$

where $\mathcal{X}_u$ denotes the set of prompts from user u, and K_x is the number of responses associated with prompt x. $\alpha_{u,\theta}^{(i,j,x)}$ represents the agreement between user u and reward model θ on the response pair (a_i, a_j) for prompt x. Z_s is the total number of such pairwise comparisons aggregated across all users in group s, A higher value of $\alpha(s)$ indicates stronger alignment between the reward model and the preferences of group s.

Agreement Disparity as Fairness Measurement. To quantify how agreement rates with reward model θ vary across demographic groups, we define the disparity measure as the maximum absolute difference between any two groups:

$$\Delta_{\text{agreement}} = \max_{s_i, s_j \in \mathcal{S}, s_i \neq s_j} |\alpha(s_i) - \alpha(s_j)| \tag{5}$$

A lower disparity value indicates that the reward model's alignment is more consistent across demographic groups, reflecting a fairer preference modeling. In contrast, a higher disparity suggests that the model preferentially aligns with certain groups, potentially perpetuating bias.

4.2 Are SOTA Reward Models Fair?

To evaluate the fairness of SOTA reward models (details in Appendix D), we compute cross-group disparities, including agreement disparity ($\Delta_{\text{agreement}}$) defined in previous section; binary and

Table 2: SOTA reward models exhibit unfairness across demographic groups. Agreement **Disparity** is quantified as the *Absolute Max-Min Deviation (%)* of agreement, as defined in Equation 5. Red shades indicate significant p-values, emphasizing that different groups demonstrate varying levels of alignment with the reward model. Full results in Appendix Table 14.

Reward Model	Size	Metric	Age	Gender	Employment Status	Education	Marital Status	English Proficiency	Religion	Ethnicity	Location
FsfairX-LLaMA3-RM-v0.1	8B	Agreement Disparity	4.4	0.8	3.6	4.3	2.0	9.0	10.0	5.8	7.3
		p-value	1.354×10^{-5}	1.682×10^{-1}	2.499×10^{-3}	2.923×10^{-3}	2.533×10^{-4}	1.296×10^{-5}	5.210×10^{-8}	9.782×10^{-12}	1.721×10^{-17}
ArmoRM-Llama3-8B-v0.1	8B	Agreement Disparity	2.0	2.3	5.3	3.2	2.9	8.4	8.4	4.5	5.6
		p-value	3.000×10^{-2}	3.606×10^{-1}	7.082×10^{-4}	9.205×10^{-1}	1.664×10^{-3}	8.767×10^{-2}	7.702×10^{-6}	1.067×10^{-6}	3.742×10^{-10}
GRM-Llama3.2-3B-rewardmodel-ft	3B	Agreement Disparity	4.6	1.1	3.5	5.7	2.3	7.4	12.7	6.0	6.2
		p-value	4.150×10^{-6}	7.225×10^{-1}	1.202×10^{-2}	7.671×10^{-1}	6.434×10^{-6}	8.750×10^{-2}	1.950×10^{-6}	6.458×10^{-8}	1.171×10^{-10}
RM-Mistral-7B	7B	Agreement Disparity	5.1	1.9	4.5	3.5	2.2	11.1	14.9	6.2	8.1
		p-value	4.276×10^{-7}	1.332×10^{-2}	3.025×10^{-4}	1.626×10^{-3}	9.218×10^{-6}	5.345×10^{-6}	8.430×10^{-10}	2.003×10^{-13}	3.503×10^{-20}
Eurus-RM-7b	7B	Agreement Disparity	3.1	2.7	4.1	5.5	5.1	6.6	9.9	4.2	5.1
		p-value	2.586×10^{-2}	1.152×10^{-1}	2.501×10^{-2}	4.818×10^{-1}	1.682×10^{-4}	3.871×10^{-1}	1.035×10^{-4}	4.813×10^{-3}	6.623×10^{-7}
RISE-Judge-Qwen2.5-7B	7B	Agreement Disparity	2.8	1.0	1.3	4.4	3.3	5.4	10.1	8.4	2.7
		p-value	1.159×10^{-2}	4.672×10^{-2}	7.065×10^{-1}	6.050×10^{-1}	2.067×10^{-1}	5.166×10^{-1}	2.168×10^{-1}	8.415×10^{-3}	2.626×10^{-1}
tulu-v2.5-13b-preference-mix-rm	13B	Agreement Disparity	5.4	5.6	3.2	6.0	3.8	6.7	14.7	6.9	6.8
		p-value	3.522×10^{-7}	4.038×10^{-5}	2.412×10^{-3}	6.381×10^{-2}	6.512×10^{-4}	1.701×10^{-4}	1.848×10^{-6}	6.726×10^{-12}	4.124×10^{-14}
LDL-Reward-Gemma-2-27B-v0.1	27B	Agreement Disparity	1.0	2.7	4.7	4.4	3.1	4.2	16.1	3.5	5.3
		p-value	7.561×10^{-1}	2.560×10^{-1}	3.570×10^{-3}	1.527×10^{-1}	7.324×10^{-2}	7.892×10^{-2}	2.351×10^{-4}	2.317×10^{-3}	9.076×10^{-6}
GRM-gemma2-2B-rewardmodel-ft	2B	Agreement Disparity	5.0	4.6	3.2	2.5	7.3	10.4	7.8	6.6	5.8
		p-value	5.462×10^{-8}	2.004×10^{-2}	5.112×10^{-3}	3.313×10^{-1}	2.034×10^{-8}	2.189×10^{-2}	3.024×10^{-4}	8.391×10^{-8}	3.823×10^{-12}

weighted F1 disparities (Δ_{F1} and $\Delta_{\text{F1-w}}$), and MCC disparity (Δ_{MCC}) similarly defined in Appendix E.1. Figure 4 visualizes the agreement score deviations from the average, with stars indicating significantly unfair treatments. Additionally, we perform Pearson's chi-squared test on the sets $\{Y_s, \hat{Y}_{s,\theta}\}_{s \in \mathcal{S}}$ to determine whether the reward model's alignment with user preferences significantly differs across groups. The results are summarized in Table 2 (full results in Table 14), where larger disparities and darker red shades indicate higher levels of unfairness. Similar visualizations but with F1 and MCC metric is illustrated with Figure 7 and Figure 8 in the Appendix.

The key takeaways from the results are as follows. (1) Reward models generally exhibit patterns of unfairness across demographic groups, with location, ethnicity, and religion emerging as the most critical factors. This may be attributed to the uneven global distribution of preference data labelers, as several regions have become prominent hubs for data labeling due to the high English proficiency and relatively low labor costs. (2) Some sensitive attributes, such as gender and education, do not exhibit as pronounced biases compared to others. This may be because preference labelers are relatively evenly distributed across these attributes. (3) The widely open-sourced LLaMA series reward models demonstrate significant unfairness overall. (4) Different preference datasets for reward modeling can lead to substantial variations in fairness, even when the base model remains the same. For instance, RM-Mistral-7B and Eurus-RM-7B show markedly different fairness levels. (5) Larger models tend to exhibit better fairness, although this observation may be confounded by the difference in preference datasets used. We leave further investigation to thoroughly examine the relationship between model scaling and fairness for future work.

5 How Reward Models Favor Group Perspectives?

While Ryan et al. (2024) observed that alignment can lead to biased improvements in language model performance across different locations, the underlying mechanisms remain unclear. In this work, we conduct a deeper analysis of the interactions between language models and reward models. Specifically, we find that reward models tend to favor perspectives from certain demographic groups over others. This phenomenon suggests that the reward models, whose preference data itself is skewed or imbalanced, may amplify and reinforce inherent biases in the language models.

Our study consists of three key steps. Given a sensitive attribute that partitions users into groups $\mathcal{S}$, we first extract group-specific perspectives from the *Controversial Questions Data* (Li et al., 2024) by explicitly prompting state-of-the-art (SOTA) language models (e.g., GPT-4o) to respond as a member of each group. Detailed information about the controversial questions data can be found in Appendix C.3. Next, we use reward models to assign a score (a.k.a., reward) to each group's response.

Table 3: Mean reward scores across demographic groups by sensitive attribute.

Sensitive Attribute	Demographic Group	Mean Reward
Age	18–24 years old	-5.4316
	25–34 years old	-5.3692
	35–44 years old	-5.3939
	45–54 years old	-5.5608
	55–64 years old	-5.6355
	65+ years old	-5.5849
	p-value	*0.0492*
Gender	Male	-5.5670
	Female	-5.4682
	Non-binary / third gender	-6.0213
	p-value	*0.0303*
Religion	Folk religion	-7.2228
	Christian	-6.9378
	Non-religious	-6.3325
	Jewish	-6.7394
	Agnostic	-6.3602
	Spiritual	-6.7769
	Muslim	-7.0822
	Hindu	-6.9106
	Buddhist	-7.2275
	Sikh	-7.0338
	p-value	*9.48e-13*
Ethnicity	White	-5.4721
	Black / African	-5.7886
	Hispanic / Latino	-5.8407
	Indigenous / First Peoples	-6.3272
	Asian	-5.8778
	Middle Eastern / Arab	-5.9991
	p-value	*3.13e-05*
Location	Northern America	-5.3430
	Africa	-5.9212
	US	-5.2954
	Latin America and the Caribbean	-6.0999
	Europe	-5.4743
	Asia	-5.4445
	Australia and New Zealand	-5.8822
	Middle East	-6.0614
	UK	-5.6950
	Oceania	-5.8178
	p-value	*4.59e-13*

Alg 1: Fair Reward Modeling (FaRM)

```
Input: Pre-trained reward model $r_\theta$,
       response pairs $\{(x_i, a_i^j, a_i^k)\}_{i \in \mathcal{D}'}$
       from different demographic groups,
       regularization coefficient $\lambda$
Output: Updated reward model $r_\theta$ with
        reduced disparity
$\theta^{(0)} \leftarrow \theta$;
for each paired response $(x_i, a_i^j, a_i^k)$ do
    $\mathcal{L}_{\text{fair}} \leftarrow \text{MSE}\left[r_\theta(x_i, a_i^j), r_\theta(x_i, a_i^k)\right]$;
    $\mathcal{L}_{\text{reg}} \leftarrow \lambda \|\theta - \theta^{(0)}\|^2$;
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{fair}} + \mathcal{L}_{\text{reg}}$;
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$;
return Updated reward model $r_\theta$
```

Finally, we compute the average score discrepancy across demographic groups to examine whether certain group perspectives are systematically preferred by the reward model.

To maintain clarity and conciseness on the main page, we present a subset of key results in Table 3, while the complete results are provided in Appendix Table 15. From the results shown in Table 3, several important patterns emerge. (1) The reward model exhibits a tendency to favor perspectives from younger individuals. (2) Perspectives from non-binary individuals are notably underrepresented, as the reward scores for non-binary perspectives are significantly lower compared to those of male and female groups. (3) On the religion axis, non-religious perspectives consistently receive higher reward scores than those of any religious group, with Muslim, folk religion, and Buddhist perspectives being treated particularly unfavorably. (4) In terms of ethnicity and location, the reward models show a stronger alignment with perspectives from North America, Europe, and Asia, as well as with white individuals. We attribute this disparity to the higher accessibility to modern language models in more developed regions, where users are more likely to interact with online LLM tools (e.g., ChatGPT), thus contributing more preference data. Additionally, the alignment with Asian perspectives may be attributed to the substantial number of data labelers from this region. Together, these findings provide a more realistic and holistic evaluation of how RLHF-aligned language models interact with reward models.

6 Mitigating Demographic Bias in Reward Models

In the previous sections, we have studied that even state-of-the-art reward models exhibit strong bias across different demographic groups partitioned by various sensitive attributes. In this work, we propose the first method, Fair Reward Modeling (FaRM) to mitigate such bias in the reward models.

6.1 Fairness Constraint with Regularization

The intuition behind our algorithm builds on the findings from Section 5, where we demonstrated that reward models exhibit preferences toward perspectives from certain demographic groups. To mitigate this bias, our algorithm explicitly targets response pairs that are treated differently by the original reward model. Specifically, we encourage the reward model to produce **equal reward scores for these paired responses**, while simultaneously applying a regularization term to maintain the model's original reward modeling capabilities.

Table 4: Performance of our FaRM. After fine-tuning, with a compromise of 8% on reward modeling capability, FaRM leads to a huge improvement on the fairness metrics.

		Fairness Metric		Reward Modeling Score			
Reward Model	Methods	Avg Agreement Disparity (↓)	Mean Square of Reward Difference (↓)	Chat (↑)	Chat Hard (↑)	Safety (↑)	Reasoning (↑)
RM-Gemma-2B	Original	0.059	0.534	94.4	40.8	49.9	76.4
	FaRM Fine-tuned	0.034	0.078	88.8	37.8	45.4	71.1
RM-LLaMA-3.2-3B	Original	0.055	2.671	91.6	84.9	92.7	94.5
	FaRM Fine-tuned	0.031	0.447	81.8	73.9	87.5	86.1

Formally, given a reward model $r_\theta(\cdot)$ with parameters θ, whose reward to a response a to prompt x is $r_\theta(x, a)$. Let $\{(x_i, a_i^1, a_i^2)\}_{i \in \mathcal{D}'}$ be response pairs from different demographic groups that exhibit significant reward disparity ($\mathcal{D}'$ denotes this synthetic dataset by us). The goal of FaRM is to minimize the difference between the predicted rewards of the pairs while preserving the model's reward modeling capability by optimizing $\mathcal{L}$:

$$\underbrace{(1-\lambda)\,\mathbb{E}[(r_\theta(x_i, a_i^1) - r_\theta(x_i, a_i^2))^2]}_{\text{Fairness Loss}} + \underbrace{\lambda\|\theta - \theta^{(0)}\|^2}_{\text{Regularization Loss}}$$

where the expectation is over $i \in \mathcal{D}'$ and λ is the regularization coefficient controlling the balance between fairness and reward modeling capability preservation.

6.2 Debiasing by FaRM

We empirically validate our FaRM algorithm by analyzing its performance on both fairness and reward modeling axes. Since the fairness loss encourages the reward models to produce similar reward values, we hypothesize that increasing the fairness loss may not necessarily enhance the reward modeling capability. In the extreme case where $\lambda = 0$, the optimal reward model that minimizes $\mathcal{L}$ would be one that outputs a constant reward value for any input, i.e., $r_\theta(x_i, a_i^1) = r_\theta(x_i, a_i^2) = c$. In contrast, if $\lambda = 1$, the optimal reward model that minimizes $\mathcal{L}$ would converge to the original unmodified reward model . Therefore, our FaRM algorithm inherently balances reward modeling capacity and fairness, akin to many existing fairness-focused machine learning approaches (Caton and Haas, 2024).

To empirically study this trade-off, we evaluate the model's performance on both the fairness and reward modeling metrics for both Gemma and LLaMA reward models. After splitting the controversial question dataset into training and testing sets, we optimize the objective function using the training data with various λ, evaluate the model on the test data and report the best-performing model. We present the results in Table 4. Average Agreement Disparity represents the average of agreement disparities across all demographic attributes[2]. Therefore, this metric ranges from 0 to 1, with smaller values indicating better fairness (i.e., less disparity across demographic groups). Our results show that FaRM could significantly improve the fairness of the reward model with a slight compromise on reward modeling capability. Specifically, we observe that model-group alignment improves for previously under-aligned groups, while it may decrease slightly for some well-aligned (often majority) groups. This trade-off can result in a modest reduction in overall model agreement, but it helps reduce demographic disparity.

7 Related Works

Due to page limit, we summary the core related works in the main page, while a more comprehensive review can be found in Appendix A. Additionally, we discuss promising future directions in Appendix B.

Reward Modeling and RLHF. The dominant RLHF approach is the three-staged PPO-based framework we have introduced in the preliminary section. Following the release of Chat-GPT, researchers have developed alternative approaches, such as iterative rejection sampling fine-tuning (Dong et al., 2023; Xiong et al., 2025) and direct alignment algorithms (Zhao et al., 2023; Rafailov et al., 2023). Although direct alignment methods were originally proposed to bypass the reward modeling stage, subsequent studies have shown that their iterative, on-policy variants outperform their offline counterparts and are now more commonly adopted in practice (Xiong et al., 2023; Xie et al., 2024). In these approaches, the intermediate policy is deployed to collect fresh responses and an external reward (preference) model is used to label them. In summary, despite methodological differences, all major RLHF and preference learning

[2] For instance, the value of 0.055 for the original RM-LLaMA-3.2-3B model corresponds to the average of its agreement disparities reported in Table 2: 0.055 = (4.6+1.1+3.5+5.7+2.3+7.4+12.7+6.0+6.2) / 9

frameworks fundamentally depend on the quality of the reward model.

Bias in Reward Models It is well known that RLHF suffers from reward hacking, where the reward model fails to fully capture true human preferences, and overfitting to this model can even degrade alignment with human intent (Gao et al., 2023). The most extensively studied case is length bias, where models exploit the reward model by generating overly long responses (Li et al., 2023a; Zheng et al., 2023; Chen et al., 2024). Subsequent works have generalized this to broader format biases, such as list-based or overly structured responses (Zhang et al., 2024a). In response, several mitigation strategies have been proposed, including reward ensembling (Coste et al., 2023), model merging (Lin et al., 2023), and correlation penalties (Chen et al., 2024). However, to the best of our knowledge, fairness issues in reward modeling, such as demographic or societal biases, remain largely under-explored.

Fairness and Inclusiveness in LLMs. The development of LLMs has motivated techniques to enhance fairness throughout the training pipeline, ranging from data augmentation and calibration (Feder et al., 2023; Stahl et al., 2022), to in-training mitigation (Thakur et al., 2023; Lauscher et al., 2021), and prompt engineering for reducing biases in outputs (Bubeck et al., 2023; Tamkin et al., 2023). Recent advancements in LLM pluralization (Sorensen et al., 2024) aim to make pre-trained models more inclusive by overtone response generation (Lake et al., 2024), steerable and personalized LLMs (Xie et al., 2025; Jiang et al., 2024), or through distributional and federated modeling (Srewa et al., 2025; Liu et al., 2020). These approaches often adopt the perspective of Language Model Agents, leveraging strong pre-trained models (Feng et al., 2024; Alamdari et al., 2024). In contrast, our work focuses on reward modeling, which has been largely overlooked by existing works.

8 Conclusion

In this work, we present a comprehensive study of fairness in reward models used for LLM alignment. Through both group preference and group perspective analyses, we demonstrate that state-of-the-art reward models exhibit significant disparities across demographic groups. To address this issue, we propose Fair Reward Modeling (FaRM), a novel method that balances fairness and reward modeling through regularization. Empirical results show that FaRM effectively mitigates demographic bias while almost maintaining model performance.

Acknowledgments

This work is supported by National Science Foundation under Award No. IIS-2117902. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

Limitations

While our study presents the most comprehensive analysis to date on fairness in reward modeling across a wide range of demographic attributes, several limitations remain. First, the analysis is constrained by the availability and representativeness of demographic annotations in publicly available datasets. As for the PRISM dataset, while we present studies in Appendix C.1 to validate the robustness, it may still underrepresent certain populations or overrepresent English-speaking and internet-connected users, potentially skewing fairness assessments. It is worth mentioning that we also cross-validate our findings of Section 4 using another dataset CIVICS (details in C.2). Second, our proposed mitigation method, FaRM, requires access to group-annotated response pairs, which may not be feasible in all deployment contexts due to privacy, availability, or annotation cost. Additionally, FaRM relies on synthetic pairing of responses across groups, which might not perfectly reflect real-world user interactions or value conflicts. Furthermore, we focus primarily on static fairness evaluation metrics such as agreement parity and average reward difference. These measures may not capture more nuanced or longitudinal effects of unfair alignment, such as enhancement of stereotypes or downstream harms in user experience.

Ethic Statement

The primary objective of this research is to introduce a series of methodologies to investigate the biases encoded in reward models, which are inherently shaped by the training data used to develop these models. By identifying and analyzing these biases, we aim to enhance the trustworthiness and ethical application of language techniques, ultimately contributing to the creation of more inclusive and fair technologies.

All datasets used in this study are publicly available and have undergone careful review to ensure they do not contain personally identifiable information (PII) or offensive content. However, given that these datasets are manually curated, there may still be limitations in terms of representativeness. As noted in the Limitations section, the sampling methodology employed during the PRISM dataset collection may not strictly adhere to ideal distribution practices due to practical constraints. Consequently, some demographic groups may be underrepresented, which could influence our evaluations.

We acknowledge the importance of critically examining these potential biases and encourage the community to interpret the results with an awareness of these limitations. Future work should aim to address these challenges by improving data collection processes and developing more comprehensive and balanced datasets.

By openly discussing these aspects, we hope to foster a more transparent and responsible approach to the study of bias in reward models and support ongoing efforts to develop AI systems that are fair, accountable, and have the potential to be aligned with diverse human values. Finally, we note that all results presented in this work are experimental and should be interpreted with caution, without overgeneralization beyond the scope of our study.

References

Parand A. Alamdari, Toryn Q. Klassen, Rodrigo Toro Icarte, and Sheila A. McIlraith. 2024. Being considerate as a pathway towards pluralistic alignment for agentic AI. *CoRR*, abs/2411.10613.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Karuna Bhaila, Minh-Hao Van, Kennedy Edemacu, Chen Zhao, Feng Chen, and Xintao Wu. 2024. Fair in-context learning via latent concept variables. *arXiv preprint arXiv:2411.02671*.

Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December*, 7.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Simon Caton and Christian Haas. 2024. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7):166:1–166:38.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Odin: Disentangled reward mitigates hacking in rlhf. *arXiv preprint arXiv:2402.07319*.

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89.

Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness in large language models: A taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1):34–48.

Michael Colacci, Yu Qing Huang, Gemma Postill, Pavel Zhelnov, Orna Fennelly, Amol Verma, Sharon Straus, and Andrea C Tricco. 2025. Sociodemographic bias in clinical machine learning models: a scoping review of algorithmic bias instances and mechanisms. *Journal of Clinical Epidemiology*, 178:111606.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*.

Mingxuan Cui, Yilan Jiang, Duo Zhou, Cheng Qian, Yuji Zhang, and Qiong Wang. 2025. Shortagesim: Simulating drug shortages under information asymmetry. *arXiv preprint arXiv:2509.01813*.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. 2024. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *Preprint*, arXiv:2406.10162.

Shizhe Diao, Rui Pan, Hanze Dong, Kashun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. 2024. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 116–127. Association for Computational Linguistics.

Thang Viet Doan, Zhibo Chu, Zichong Wang, and Wenbin Zhang. 2024. Fairness definitions in language models explained. *CoRR*, abs/2407.18454.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. RLHF workflow: From reward modeling to online RLHF. *CoRR*, abs/2405.07863.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *CoRR*, abs/2306.16388.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Zahra Fatemi, Chen Xing, Wenhao Liu, and Caiming Xiong. 2023. Improving gender fairness of pretrained language models without catastrophic forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1249–1262. Association for Computational Linguistics.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. 2023. Data augmentations for improved (large) language model generalization. *Advances in Neural Information Processing Systems*, 36:70638–70653.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4151–4171. Association for Computational Linguistics.

Gillian Franklin, Rachel Stephens, Muhammad Piracha, Shmuel Tiosano, Frank Lehouillier, Ross Koppel, and Peter L Elkin. 2024. The sociodemographic biases in machine learning algorithms: a biomedical informatics perspective. *Life*, 14(6):652.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for RLHF. *CoRR*, abs/2410.14872.

Dongqi Fu, Liri Fang, Zihao Li, Hanghang Tong, Vetle I. Torvik, and Jingrui He. 2024. Parametric graph representations in the era of foundation models: A survey and position. *CoRR*, abs/2410.12126.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Pratyush Garg, John D. Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, pages 3662–3666. IEEE.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5448–5458. Association for Computational Linguistics.

Zara Hall, Melanie Subbiah, Thomas P. Zollo, Kathleen McKeown, and Richard S. Zemel. 2025. Guiding LLM decision-making with fairness reward models. *CoRR*, abs/2507.11344.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Jingyu Hu, Weiru Liu, and Mengnan Du. 2024. Strategic demonstration selection for improved fairness in llm in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7460–7475.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. 2024. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? *CoRR*, abs/2410.03868.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *CoRR*, abs/2312.14925.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew M. Bean, Katerina Margatina, Rafael Mosquera Gómez, Juan Ciro, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott Hale. 2024. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Joonhyuk Ko, Juba Ziani, Saswat Das, Matt Williams, and Ferdinando Fioretto. 2025. Fairness issues and mitigations in (differentially private) sociodemographic data processes. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 28160–28167. AAAI Press.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference, CI 2023, Delft, Netherlands, November 6-9, 2023*, pages 12–24. ACM.

Thom Lake, Eunsol Choi, and Greg Durrett. 2024. From distributional to overton pluralism: Investigating large language model alignment. *CoRR*, abs/2406.17692.

Anne Lauscher, Tobias Lüken, and Goran Glavas. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4782–4797. Association for Computational Linguistics.

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024. Can llms speak for diverse people? tuning llms via debate to generate controllable controversial statements. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 16160–16176. Association for Computational Linguistics.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023a. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023b. A survey on fairness in large language models. *CoRR*, abs/2308.10149.

Zihao Li, Lecheng Zheng, Bowen Jin, Dongqi Fu, Baoyu Jing, Yikun Ban, Jingrui He, and Jiawei Han. 2025. Can graph neural networks learn language with extremely weak text supervision? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11138–11165. Association for Computational Linguistics.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and 1 others. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*, 11:14.

Tianci Liu, Haoyu Wang, Shiyang Wang, Yu Cheng, and Jing Gao. 2024a. Lidao: towards limited interventions for debiasing (large) language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 32083–32099.

Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. 2020. Self-paced ensemble for highly imbalanced massive data classification. In *2020 IEEE 36th international conference on data engineering (ICDE)*, pages 841–852. IEEE.

Zhining Liu, Zihao Li, Ze Yang, Tianxin Wei, Jian Kang, Yada Zhu, Hendrik F. Hamann, Jingrui He, and Hanghang Tong. 2025. CLIMB: class-imbalanced learning benchmark on tabular data. *CoRR*, abs/2505.17451.

Zhining Liu, Ruizhong Qiu, Zhichen Zeng, Yada Zhu, Hendrik Hamann, and Hanghang Tong. 2024b. Aim: Attributing, interpreting, mitigating data unfairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2014–2025.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35.

Michele Merler, Nalini K. Ratha, Rogério Schmidt Feris, and John R. Smith. 2019. Diversity in faces. *CoRR*, abs/1901.10436.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against african american language in reward models. *CoRR*, abs/2502.12858.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Ki-Woong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xuying Ning, Dongqi Fu, Tianxin Wei, Wujiang Xu, and Jingrui He. 2025. Graph4MM: Weaving multimodal learning with structural information. In *Forty-second International Conference on Machine Learning*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sheng Ouyang, Yulan Hu, Ge Chen, Qingyang Li, Fuzheng Zhang, and Yong Liu. 2025. Towards reward fairness in RLHF: from a resource allocation perspective. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3247–3259. Association for Computational Linguistics.

Giada Pistilli, Alina Leidinger, Yacine Jernite, Atoosa Kasirzadeh, Alexandra Sasha Luccioni, and Margaret Mitchell. 2024. CIVICS: building a dataset for examining culturally-informed values in large language models. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24) - Full Archival Papers, October 21-23, 2024, San Jose, California, USA - Volume 1*, pages 1132–1144. AAAI Press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024. Warp: On the benefits of weight averaged rewarded policies. *arXiv preprint arXiv:2406.16768*.

Michael J. Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 16121–16140. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Kefan Song, Jin Yao, Runnan Jiang, Rohan Chandra, and Shangtong Zhang. 2025. Towards large language models that benefit for all: Benchmarking group fairness in reward models. *CoRR*, abs/2503.07806.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Mahmoud Srewa, Tianyu Zhao, and Salma Elmalaki. 2025. Plurallm: Pluralistic alignment in llms via federated learning. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems, HumanSys 2025, Irvine, CA, USA, May 6-9, 2025*, pages 64–69. ACM.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. 2022. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pages 39–51.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *CoRR*, abs/2312.03689.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14,* *2023*, pages 340–351. Association for Computational Linguistics.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, and 7 others. 2025. A survey on post-training of large language models. *CoRR*, abs/2503.06072.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Caroline Wang, Bin Han, Bhrij Patel, Feroze Mohideen, and Cynthia Rudin. 2020. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *CoRR*, abs/2005.04176.

Haonan Wang, Ziwei Wu, and Jingrui He. 2024a. Fairif: Boosting fairness in deep learning via influence functions with validation set sensitive attributes. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining, WSDM 2024, Merida, Mexico, March 4-8, 2024*, pages 721–730. ACM.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Tianxin Wei and Jingrui He. 2022. Comprehensive fair meta-learned recommender system. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 1989–1999. ACM.

Tianxin Wei, Bowen Jin, Ruirui Li, Hansi Zeng, Zhengyang Wang, Jianhui Sun, Qingyu Yin, Hanqing Lu, Suhang Wang, Jingrui He, and Xianfeng Tang. 2024. Towards unified multi-modal personalization: Large vision-language models for generative recommendation and beyond. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Ziwei Wu and Jingrui He. 2022. Fairness-aware model-agnostic positive and unlabeled learning. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*, pages 1698–1708. ACM.

Ziwei Wu, Lecheng Zheng, Yuancheng Yu, Ruizhong Qiu, John R. Birge, and Jingrui He. 2024. Fair anomaly detection for imbalanced groups. *CoRR*, abs/2409.10951.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. 2024. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, and 1 others. 2025. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2023. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*.

Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and 1 others. 2025. A minimalist approach to llm reasoning: from rejection sampling to reinforce. *arXiv preprint arXiv:2504.11343*.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2024a. A study of implicit ranking unfairness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7957–7970.

Rongwu Xu, Xuan Qi, Zehan Qi, Wei Xu, and Zhijiang Guo. 2024b. Debateqa: Evaluating question answering on debatable knowledge. *CoRR*, abs/2408.01419.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. In *Advances in Neural Information Processing Systems*.

Mohammad Yaseliani, Md Noor-E-Alam, and Md Mahmudul Hasan. 2024. Mitigating sociodemographic bias in opioid use disorder prediction: Fairness-aware machine learning framework. *Jmir ai*, 3:e55820.

Jiachen Yu, Shaoning Sun, Xiaohui Hu, Jiaxu Yan, Kaidong Yu, and Xuelong Li. 2025. Improve llm-as-a-judge ability as a general ability. *Preprint*, arXiv:2502.11689.

Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. 2023. Mixup-based unified framework to overcome gender bias resurgence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1755–1759. ACM.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Xuanchang Zhang, Wei Xiong, Lichang Chen, Tianyi Zhou, Heng Huang, and Tong Zhang. 2024a. From lists to emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704*.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R Fung, Jing Li, Manling Li, and Heng Ji. 2024b. Knowledge overshadowing causes amalgamated hallucination in large language models. *arXiv preprint arXiv:2407.08039*.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, ChengXiang Zhai, Manling Li, and Heng Ji. 2025. The law of knowledge overshadowing: Towards understanding, predicting and preventing LLM hallucination. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23340–23358. Association for Computational Linguistics.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Lecheng Zheng, Baoyu Jing, Zihao Li, Hanghang Tong, and Jingrui He. 2024. Heterogeneous contrastive learning for foundation models and beyond. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6666–6676. ACM.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.

Jiaru Zou, Yikun Ban, Zihao Li, Yunzhe Qi, Ruizhong Qiu, Ling Yang, and Jingrui He. 2025a. Transformer copilot: Learning from the mistake log in LLM finetuning. *CoRR*, abs/2505.16270.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2025b. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. *CoRR*, abs/2506.18896.

A More Related Works and Discussions

Before turning to more related works, we want to position this work more clearly by discussing **key difference between fairness in reward modeling and classical fairness tasks.**

- **Demographic Diversity Challenge.** Different demographic groups may hold conflicting views. As we analyzed in Section 4, the preferences of different demographic groups can vary. Given that the RLHF data labeler may come from any demographic group, this poses a root challenge of why we cannot achieve accuracy and fairness at the same time.
- **Data Distribution Constraints.** The RLHF dataset may be very skewed because the label distribution is affected by many real-world constraints. We analyzed this issue by comparing the PRISM data distribution and the distribution from reward models learned from RLHF datasets. And, starting from the pretrained reward model, we don't have full control of the fairness mitigation process, compared to classical fairness tasks where we could train a classifier from scratch.
- **Sensitive Attribute Uncertainty in RLHF data.** In real RLHF datasets, we are unable to know which sensitive attribute a data sample is associated with, compared to the classical fairness tasks, where we have access to the sensitive attribute of each data sample.

Bias in reward models has important implications for different sociodemographic groups, as disparities in alignment can systematically privilege some perspectives while marginalizing others (Colacci et al., 2025; Yaseliani et al., 2024; Franklin et al., 2024). Such disparities risk reinforcing existing societal inequities, as models aligned through biased reward signals may produce outputs that underrepresent or mischaracterize certain groups (Ko et al., 2025). Similar unfairness propagation has been observed in various domains, including recommendation (Wei and He, 2022), hiring decisions (Bogen and Rieke, 2018), identity verification (Merler et al., 2019) and criminal justice analysis (Wang et al., 2020), where machine learning technologies have been applied. Addressing these concerns is thus essential not only for technical fairness but also for ensuring that LLMs serve diverse populations equitably.

Reward Modeling and RLHF. In this work, we explore the fairness issue of reward models within the framework of Reinforcement Learning from Human Feedback (RLHF). The standard RLHF pipeline in the application of Chat-GPT and Claude (Ouyang et al., 2022; Bai et al., 2022) consists of three stages: human preference data collection, reward modeling from the collected preference data, and model optimization from the learned reward. In particular, we typically optimize the model using Proximal Policy Optimization (PPO) (Schulman et al., 2017), hence the name RLHF. However, the RL training with PPO is rather computationally expansive and unstable. In recognition of this, researchers have developed alternative approaches to simplify the RLHF process. These approaches include the iterative rejection sampling fine-tuning (Dong et al., 2023; Xiong et al., 2025; Touvron et al., 2023; Zou et al., 2025b), and the direct alignment algorithms (Zhao et al., 2023; Rafailov et al., 2023). Rejection sampling fine-tuning iteratively deploys the current LLMs to collect a large amount of responses per prompt, uses a reward model to filter out the low-quality responses, and then finetune the model on the rest of the data. In contrast, direct alignment methods were originally proposed to bypass the reward modeling stage. However, subsequent studies have shown that their iterative, on-policy variants outperform their offline counterparts and are now more commonly adopted in practice (Xiong et al., 2023; Xie et al., 2024). In these iterative variants, the intermediate policy is deployed to collect fresh responses and an external reward (preference) model is used to label them. In summary, despite methodological differences, all major RLHF and preference learning frameworks fundamentally depend on the quality of the reward model. Therefore, the fairness bias in the reward model (and the preference dataset) will inevitably transfer to the downstream aligned LLMs.

Bias in Reward Models It is well known that RLHF suffers from reward hacking, where the reward model fails to fully capture true human preferences. Overfitting to this model can even degrade alignment with human intent (Gao et al., 2023; Denison et al., 2024) because it is usually much easier to exploit these bias patterns to chase for a high reward compared to improving the cotent quality. The most extensively studied case is length bias, where models exploit the reward model and preference model (e.g., human and GPT-4) by generating overly long responses (Li et al., 2023a; Chen et al., 2024; Zheng et al., 2023). Subsequent works have generalized this to broader format biases, such as list-based or overly structured responses (Zhang et al., 2024a). These biases are also exploited by the current LLMs to chase for a high ranking in the benchmark leaderboard (Zheng et al., 2023). In response, several mitigation strategies have been proposed for the RLHF training, including reward ensembling (Coste et al., 2023), model merging (Lin et al., 2023; Ramé et al., 2024), and correlation penalties (Chen et al., 2024). Moreover, Li et al. (2023a) and Zheng et al. (2023) have also introduced length-controlled and format-controlled version of their benchmarks, aiming for mitigating the impacts of these bias patterns. However, to the best of our knowledge, fairness issues in reward modeling, such as demographic or societal biases, remain largely under-explored.

Fairness and Inclusiveness in LLMs. As LLMs are increasingly used in high-stakes applications (Zou et al., 2025a; Fu et al., 2024), there has been a surge of interest in examining the fairness and inclusiveness in LLMs (Gallegos et al., 2024; Chu et al., 2024). A wide range of techniques have been explored to detect and enhance fairness throughout the training pipeline. Data-level interventions include counterfactual data augmentation and calibration (Feder et al., 2023; Stahl et al., 2022; Xu et al., 2024a; Wu and He, 2022), data curation (Hu et al., 2024) and strategic demonstration selection (Bhaila et al., 2024). These methods improve exposure to marginalized identities but often struggle to scale or address more nuanced or intersectional implicit bias. To complement data-level fixes, in-processing techniques have been proposed including adding fairness constraints (Liu et al., 2024a; Wang et al., 2024a), few-shot debiasing (Thakur et al., 2023) and modular adaption (Lauscher et al., 2021; Wu et al., 2024). Prompt-based strategies have also gained traction in reducing biases in outputs (Bubeck et al., 2023; Tamkin et al., 2023). Recent advancements in LLM pluralization (Sorensen et al., 2024) aim to make pre-trained models more inclusive by overtone response generation (Lake et al., 2024), steerable and personalized LLMs (Xie et al., 2025; Jiang et al., 2024; Wei et al., 2024), or through distributional and federated modeling (Srewa et al., 2025). These approaches often adopt the perspective of Language Model Agents, leveraging strong pre-trained models (Feng et al., 2024; Alamdari et al., 2024; Ning et al., 2025). In contrast, our work focuses on reward modeling, which has been largely overlooked by existing works.

B Future Directions

While this work presents the first comprehensive analysis of fairness in reward models and introduces FaRM as an effective mitigation approach, we list several open questions for future exploration.

Generalization to New or Multi-Valued Groups. Demographic groups are finite and enumerable in practice. This means that preference data can be collected to cover a broad set of groups. Beyond this, FaRM is naturally extensible to unseen or multi-valued demographic groups through natural language descriptions of group identities. Since our method synthesizes group-specific perspectives via prompting, it has the potential to generalize to novel contexts (Zheng et al., 2024; Li et al., 2025) by conditioning on descriptive attributes rather than predefined categories.

Improved Data Collection. Our analysis is currently constrained by existing datasets such as PRISM and CIVICS. Future work should investigate methods to collect more balanced and representative preference datasets. This includes participatory data collection, federated protocols, and incentivizing underrepresented groups to ensure that fairness assessments reflect a truly global population.

Integration with Alignment Methods. Although our study primarily focuses on fairness in reward modeling, modern RLHF pipelines ultimately align large language models with the outputs of these trained reward models. This raises an important open question: to what extent do improvements in reward model fairness translate into fairer behavior in the final aligned LLMs? A very recent work (Hall et al., 2025) made an initial attempt to bridge the gap between a pre-trained fair reward model and fair language modeling, and we expect more future works in this direction conducting end-to-end evaluations.

Better RM Bias Mitigation. FaRM is intended as a first step toward addressing bias in reward modeling. Rather than a definitive solution, we position it as a strong and interpretable baseline that highlights the core challenge of reward model fairness, a space largely overlooked compared to LLM fairness. By providing a concrete starting point, FaRM motivates future work on developing and integrating more advanced mitigation strategies tailored to reward model training.

Socialdemographic studies. From a sociological perspective, an important future direction is to examine how algorithmic unfairness in LLMs manifests in the lived experiences of different demographic groups (Cui et al., 2025). While our work reveals disparities at the level of reward modeling, the broader question concerns how such biases propagate into real-world outcomes, such as unequal access to information, differential treatment in customer service, or reinforcement of stereotypes in educational and workplace settings.

C Dataset Details

C.1 The PRISM Alignment Dataset

The PRISM Alignment Dataset[3] (Kirk et al., 2024) is a comprehensive resource designed to enhance the alignment of Large Language Models (LLMs) with diverse human values and preferences. Developed by researchers from the University of Oxford and collaborators, PRISM emphasizes participatory, representative, and individualized human feedback to address the complexities of subjective and multicultural alignment in AI systems.

PRISM comprises 8,011 multi-turn conversations between 1,500 participants from 75 countries and 21 different LLMs. Each participant completed a detailed survey capturing sociodemographic information and stated preferences before engaging in real-time interactions with LLMs. These interactions were followed by fine-grained evaluations, including ratings on attributes such as factuality, fluency, and creativity, as well as open-ended feedback.

As outlined in Section 3, we present an overview of the data in Tables 5 and 6, along with the statistics in Table 9.

Correlation Study of Demographic Group Attributes. The primary goal of this correlation study is to validate the PRISM dataset as a robust and representative resource for analyzing human preferences. A high-quality dataset should naturally reflect correlations among demographic attributes, and verifying these relationships is essential to assess the dataset's validity. For instance, individuals aged 65 or older are more likely to be retired, and individuals from the Africa location are more likely to be Black/African people.

To achieve this, we compute the average preference vector across the nine preference axes for each demographic group and use these vectors as features to construct a correlation matrix. The resulting heatmap, shown in Figure 5, illustrates how demographic attributes correlate with one another, helping to ensure that the PRISM dataset accurately captures realistic demographic interactions.

The heatmap reveals that while specific correlation values vary between demographic attributes, the columns and rows corresponding to religion, ethnicity, and location exhibit a more sparse pattern. This indicates that these attributes tend to partition individuals into demographic groups with greater preference discrepancies. These results align with our findings in Section 4 and Section 5, where we observed that the reward model struggles to capture preference variations among groups distinguished by certain sensitive attributes.

By confirming the presence of realistic correlations and preference discrepancies among demographic groups, this study demonstrates that the PRISM dataset accurately captures human diversity and variability, affirming its validity as a valuable resource for reward model evaluation.

C.2 The CIVICS Culturally-Informed Values Dataset

The CIVICS dataset is a multilingual, annotated collection of civic discourse statements sourced from various countries and institutional contexts. Each entry includes the original statement, its English translation when applicable, a topical category (e.g., disability rights), the issuing organization, and source documentation. Crucially, each statement is annotated by multiple human annotators (typically three), who label the underlying civic values conveyed—such as equality, support, or accessibility. These annotations capture both consensus and disagreement, offering a nuanced view of how public values are communicated. The dataset is well-suited for cross-linguistic and cross-cultural studies of value expression in public communication. Refer to Table 8 for examples.

Table 7 presents the mean reward scores across different geographic locations in the CIVICS dataset. This analysis serves as a validation step to assess whether the patterns observed in Section 4 using the PRISM dataset are consistent across independent datasets. Notably, we again observe a strong regional skew: responses from European countries (e.g., Germany, Italy, France)

[3] https://hannahkirk.github.io/prism-alignment

Table 5: Demographic information for example users from the PRISM dataset. *Note: Each row represents a separate user.*

User ID	Age	Gender	Employment	Education	Marital Status	English Proficiency	Religion
user0	35–44 years old	Male	Working full-time	University Bachelors Degree	Never been married	Native speaker	Prefer not to say
user1	35–44 years old	Male	Working full-time	Some Secondary	Never been married	Native speaker	Prefer not to say
user2	18–24 years old	Female	Working full-time	University Bachelors Degree	Married	Fluent	Folk religion

Table 6: Stated preferences for example PRISM users. *Note: Each row represents a separate user. Scores are computed per user.*

User ID	Stated Preference (Score)							
	Values	Creativity	Fluency	Factuality	Diversity	Safety	Personalisation	Helpfulness
user0	83	100	100	100	100	90	71	100
user1	19	73	86	79	83	100	84	100
user2	100	100	100	100	99	100	75	100

Table 7: Mean reward scores across different locations for English data in the CIVICS dataset, with a Kruskal-Wallis p-value of 4.443×10^{-9}.

Location	Mean Reward
Australia	0.648
Canada	0.704
Singapore	1.559
France	1.683
Germany	1.832
Italy	2.241
Turkey	1.481

tend to receive higher reward scores than those from countries like Australia or Canada. This supports the finding that reward models exhibit geographic preferences, suggesting that the biases identified in PRISM are not dataset-specific but rather reflect broader tendencies in current reward modeling practices. The Kruskal-Wallis p-value (4.443×10^{-9}) further confirms that the differences across locations are statistically significant.

C.3 The DEBATunE Controversial Question Dataset

we use the DEBATunE dataset (Li et al., 2024) as part of the prompts for generating group perspectives as shown in Figure 2 (right). The DEBATunE dataset is a comprehensive resource designed to enhance the controllability of Large Language Models. It comprises 710 controversial topics, covering a wide range of societal, ethical, and political issues, providing a rich ground for studying LLM behavior in contentious contexts.

Sample Topics:

- Should internet gambling be legalized?
- Should we have comprehensive sex education in schools?
- Is AI replacing human jobs?
- Should we end racial profiling?
- Should we ban abortions?

To generate group perspectives, given a `question` from the DEBATunE dataset, for a specific demographic group with `unique_attribute` for `sensitive_attribute`, we used the following prompt:

> **Prompt:** For sensitive attribute `{sensitive_attribute}`, imagine you are `{unique_attribute}`. Limit your response to 50 words. How would you answer: `{question}`?

ID	Country	Topic	Statement	Translation	Org	Source	Final Label	Annotators (1–3)
37	France	Disability Rights	L'inclusion des personnes en situation de handicap constitue une des priorités du quinquennat.	The inclusion of people with disabilities is one of the priorities of the five-year term.	Gouvernement (France)	French Government	equality	• equality • equality • equality
60	Germany	Disability Rights	Wir setzen uns das Ziel, alle öffentlichen Gebäude barrierefrei zu gestalten.	We have set ourselves the goal of making all federal public buildings fully barrier-free.	Bundesregierung	Coalition Agreement 2021	accessibility, support	• accessibility, support • accessibility, support • accessibility, support
1	Australia	Disability Rights	Programs that help people with disability to maintain their housing tenancy can lead to more secure long-term housing arrangements and greater independence.		AIHW	Australian Institute of Health and Welfare	support	• support • support

Table 8: Three examples from CIVICS Dataset. The third annotator was unavailable English samples.

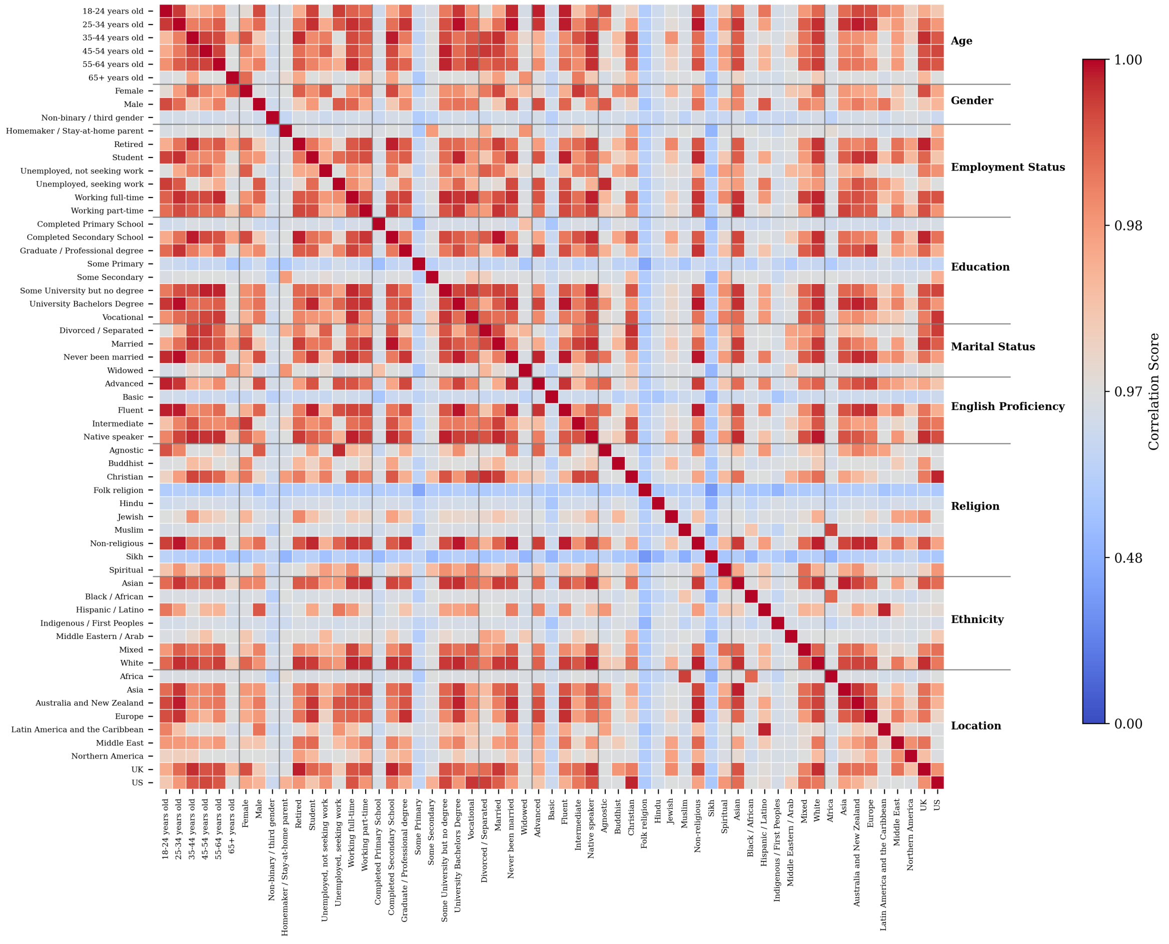


Figure 5: Correlation heatmap of demographic group attributes in the PRISM dataset. The results demonstrate that the PRISM dataset is robust and representative. Given the general similarity of human preferences, we experimented with multiple thresholds and found that a correlation score of 0.97 is good to be considered significant. Attributes related to religion, ethnicity, and location exhibit more sparse correlations, indicating greater variability in preferences within these groups, which is consistent with our findings in Section 4 and Section 5.

Table 9: Summary statistics by group across sensitive attributes in the PRISM dataset. Dominant conversation types are indicated as Controversy Guided (†), Values Guided (‡), or Unguided (no marker).

Sensitive Attribute	Group Name	# Interactions	# Users	Avg Score	% Chosen	Dominant Type	Avg Turns per Interaction
Age	18-24 years old	5374	280	64.62	39.89%	37.68%	2.52
	25-34 years old	8562	427	65.23	40.38%	38.29%	2.50
	35-44 years old	4486	224	64.74	40.45%	41.88%	2.50
	45-54 years old	3669	189	65.05	40.16%	39.72%	2.52
	55-64 years old	3367	181	66.70	39.69%	43.30%	2.55
	65+ years old	1714	95	64.38	39.59%	39.84%	2.56
Gender	Female	13001	677	66.68	40.08%	39.88%	2.52
	Male	13705	697	63.68	40.15%	39.43%	2.51
	Non-binary / third gender	448	20	65.36	41.09%	40.09%	2.47
	Prefer not to say[*]	18	2	49.00	40.91%	81.82%	2.44
Employment Status	Homemaker / Stay-at-home parent	830	44	67.69	40.09%	39.61%	2.52
	Prefer not to say[*]	385	18	64.03	40.82%	41.45%	2.48
	Retired	1793	95	63.39	39.86%	43.03%	2.54
	Student	3412	178	65.28	40.17%	36.69%	2.51
	Unemployed, not seeking work	885	43	72.50	40.55%	40.55%	2.49
	Unemployed, seeking work	2252	108	65.22	40.82%	39.64%	2.48
	Working full-time	12899	664	64.60	40.06%	39.85%	2.52
	Working part-time	4716	246	65.39	39.93%	39.84%	2.52
Education	Completed Primary School	254	14	68.27	39.23%	35.54%†	2.56
	Completed Secondary School	3712	191	66.38	40.08%	42.84%	2.52
	Graduate / Professional degree	4531	227	63.34	40.18%	38.99%	2.52
	Prefer not to say[*]	118	7	63.88	40.20%	45.51%	2.55
	Some Primary	65	3	68.15	40.12%	79.01%	2.49
	Some Secondary	399	23	66.94	39.31%	38.43%	2.56
	Some University but no degree	4276	215	65.28	40.36%	37.76%	2.50
	University Bachelors Degree	11678	605	65.08	40.15%	39.61%	2.51
	Vocational	2139	111	66.02	39.81%	39.15%	2.54
Marital Status	Divorced / Separated	2116	109	65.60	40.07%	38.72%	2.52
	Married	8102	429	65.76	39.94%	41.32%	2.53
	Never been married	16189	819	64.73	40.23%	38.83%	2.51
	Prefer not to say[*]	374	18	66.45	40.72%	45.93%	2.46
	Widowed	391	21	64.85	39.84%	40.34%	2.53
English Proficiency	Advanced	2993	151	66.13	40.44%	39.69%	2.50
	Basic	93	5	67.11	40.34%	57.51%	2.51
	Fluent	7306	380	64.87	40.03%	38.16%	2.52
	Intermediate	771	40	68.50	40.12%	43.53%	2.51
	Native speaker	16009	820	64.89	40.12%	40.09%	2.52
Religion	Agnostic	1310	63	64.30	40.14%	37.92%	2.50
	Buddhist	165	10	65.91	39.34%	35.31%	2.56
	Christian	8600	458	67.17	39.89%	39.86%	2.53
	Folk religion	143	6	62.27	43.02%	46.80%	2.41
	Hindu	98	5	56.34	39.68%	42.91%	2.52
	Jewish	700	35	62.14	40.39%	41.70%	2.51
	Muslim	591	29	65.89	40.52%	41.39%	2.53
	Non-religious	14049	711	64.27	40.25%	39.61%	2.51
	Other[*]	52	3	53.25	40.60%	39.10%‡	2.56
	Prefer not to say[*]	1135	57	63.94	40.28%	38.48%	2.50
	Sikh	52	3	65.64	37.68%	39.86%†	2.65
	Spiritual	277	16	66.16	39.08%	43.14%	2.58
Ethnicity	Asian	1756	91	63.68	40.12%	42.34%	2.52
	Black / African	2020	117	70.26	39.08%	40.03%	2.59
	Hispanic / Latino	2359	117	65.85	40.35%	38.20%	2.50
	Indigenous / First Peoples	148	8	70.59	40.00%	36.80%	2.53
	Middle Eastern / Arab	280	12	62.80	41.32%	37.94%	2.43
	Mixed	1405	67	65.83	40.97%	37.90%	2.46
	Other[*]	289	13	56.84	40.70%	36.36%	2.47
	Prefer not to say[*]	1333	79	64.48	39.93%	41.88%	2.53
	White	17582	892	64.69	40.15%	39.65%	2.52
Location	Africa	1970	114	71.26	39.00%	39.66%	2.59
	Asia	1162	57	63.14	40.76%	44.30%	2.48
	Australia and New Zealand	2433	122	64.17	40.23%	37.75%	2.50
	Europe	5824	292	64.25	40.34%	38.63%	2.51
	Latin America and the Caribbean	2837	143	66.14	40.35%	37.65%	2.49
	Middle East	826	43	60.79	39.94%	42.97%	2.52
	Northern America	949	48	64.65	40.53%	43.31%	2.50
	Oceania[*]	28	1	84.56	41.18%	55.88%‡	2.43
	Prefer not to say[*]	54	3	52.30	39.13%	52.17%	2.56
	UK	5277	271	64.37	39.88%	38.93%	2.53
	US	5812	302	65.60	40.24%	41.25%	2.51

[*] Groups labeled "Prefer not to say", "Other", and "Oceania" are excluded from statistical comparisons because they either aggregate heterogeneous populations or have insufficient users for reliable p-value estimation.

D Reward Model Details

In this section, we detail the reward models examined in this study. The selected models represent a diverse range of architectures, training methodologies, and data sources, enabling a comprehensive analysis of state-of-the-art approaches to reward modeling. A summary of the reward models we used is provided in Table 10.

GRM-Llama3.2-3B-rewardmodel-ft. The GRM-Llama3.2-3B-rewardmodel-ft[4] model, introduced by Yang et al. (2024), is built on the Meta-Llama-3.2-3B-Instruct base model. It has approximately 3 billion parameters and is fine-tuned using the Skywork-Reward-Preference-80K-v0.2 dataset. The model employs the Bradley–Terry ranking method, which is well-suited for comparative preference learning. This model is designed to leverage robust human feedback data, making it suitable for alignment tasks.

ArmoRM-Llama3-8B-v0.1. The ArmoRM-Llama3-8B-v0.1[5], proposed by Wang et al. (2024b), uses the Meta-Llama-3-8B-Instruct model as a base, featuring 8 billion parameters. It utilizes a mixture of datasets including HelpSteer, UltraFeedback, BeaverTails-30k, and others. The model's unique training strategy, combining ArmoRM with Mixture of Experts (MoE), aims to enhance interpretability and adaptability across diverse preference contexts.

FsfairX-LLaMA3-RM-v0.1. Presented by Dong et al. (2024), FsfairX-LLaMA3-RM-v0.1 shares the same base model as ArmoRM but incorporates a different dataset mixture, including HH-RLHF, SHP, and PKU-SafeRLHF-30K. The model follows the Bradley–Terry method, emphasizing fairness and robustness through data diversity.

tulu-v2.5-13b-preference-mix-rm. The tulu-v2.5-13b-preference-mix-rm[6] model, from Ivison et al. (2024), is based on the Llama-2-13b-hf and trained with the Tulu 2.5 preference mix. It utilizes Proximal Policy Optimization (PPO) for reward modeling, focusing on fine-tuning large language models using reinforcement learning techniques to better capture nuanced preferences.

GRM-gemma2-2B-rewardmodel-ft. This GRM-gemma2-2B-rewardmodel-ft[7] model, also introduced by Yang et al. (2024), is a smaller variant compared to its Llama-based counterpart, featuring 2 billion parameters. It is derived from Gemma-2-2B and trained on the Skywork-Reward-Preference-80K-v0.2 dataset. The combined use of Supervised Fine-Tuning (SFT) and Bradley–Terry enhances its ability to model human preferences accurately while maintaining model efficiency.

LDL-Reward-Gemma-2-27B-v0.1. The LDL-Reward-Gemma-2-27B-v0.1[8] model is the largest among those analyzed, with 27 billion parameters. It is based on Gemma-2-27B and trained using a combination of label distillation and Bradley–Terry methods. While the training dataset is not explicitly specified, the large parameter count suggests it is optimized for capturing complex reward functions.

RM-Mistral-7B. RM-Mistral-7B[9], introduced by Xiong et al. (2023), is based on the Mistral-7B-Instruct-v0.2 model. It is trained on a mixture of datasets, including HH-RLHF, SHP, and Ultra-Feedback. The use of Bradley–Terry ranking ensures reliable modeling of human feedback, making it effective for multi-context preference aggregation.

Eurus-RM-7b. Proposed by Yuan et al. (2024), Eurus-RM-7b[10] shares the Mistral-7B-Instruct-v0.2 base model but uses a different dataset mixture, including UltraInteract, UltraFeedback,

[4] https://huggingface.co/Ray2333/GRM-Llama3.2-3B-rewardmodel-ft

[5] https://huggingface.co/RLHFlow/ArmoRM-Llama3-8B-v0.1

[6] https://huggingface.co/sfairXC/FsfairX-LLaMA3-RM-v0.1

[7] https://huggingface.co/Ray2333/GRM-gemma2-2B-rewardmodel-ft

[8] https://huggingface.co/ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1

[9] https://huggingface.co/weqweasdas/RM-Mistral-7B

[10] https://huggingface.co/openbmb/Eurus-RM-7b

Table 10: Details of the reward models examined in this work. We conduct a comprehensive study of state-of-the-art reward models, covering a diverse set of base language models, training datasets, and training methodologies.

Model Name	Size	Base Model	Training Dataset(s)	Training Method
GRM-Llama3.2-3B-rewardmodel-ft (Yang et al., 2024)	3B	Meta-Llama-3.2-3B-Instruct	Skywork-Reward-Preference-80K-v0.2	Bradley–Terry
ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024b)	8B	Meta-Llama-3-8B-Instruct	Mixture of datasets[1]	ArmoRM + MoE
FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2024)	8B	Meta-Llama-3-8B-Instruct	Mixture of datasets[2]	Bradley–Terry
tulu-v2.5-13b-preference-mix-rm (Ivison et al., 2024)	13B	Llama-2-13b-hf	Tulu 2.5 preference mix	PPO
GRM-gemma2-2B-rewardmodel-ft (Yang et al., 2024)	2B	Gemma-2-2B	Skywork-Reward-Preference-80K-v0.2	SFT + Bradley–Terry
LDL-Reward-Gemma-2-27B-v0.1	27B	Gemma-2-27B	Not specified	label distillation + Bradley–Terry
RM-Mistral-7B (Xiong et al., 2023)	7B	Mistral-7B-Instruct-v0.2	Mixture of datasets[3]	Bradley–Terry
Eurus-RM-7b (Yuan et al., 2024)	7B	Mistral-7B-Instruct-v0.2	Mixture of datasets[4]	Bradley–Terry
RISE-Judge-Qwen2.5-7B (Yu et al., 2025)	7B	Qwen2.5-7B-Base	Mixture of datasets[5]	SFT + DPO

[1] Includes HelpSteer, UltraFeedback, BeaverTails-30k, CodeUltraFeedback, Prometheus, Argilla-Capybara, Argilla-OpenOrca, Argilla-Math-Preference.
[2] Includes HH-RLHF, SHP, HelpSteer, PKU-SafeRLHF-30K, UltraFeedback, UltraInteract, CodeUltraFeedback, Argilla-Math, OpenOrca, Capybara.
[3] Includes HH-RLHF, SHP, Ultra-Feedback, Capybara, HelpSteer, Orca, and PKU-SafeRLHF-30K.
[4] Includes UltraInteract, UltraFeedback, UltraSafety.
[5] Includes Math-PRM800K Dataset, Skywork-Reward-Preference-80K-v0.2, Non-Judge Dataset.

and UltraSafety. This model also leverages the Bradley–Terry approach, emphasizing balanced and safe reward modeling.

RISE-Judge-Qwen2.5-7B The RISE-Judge-Qwen2.5-7B[11] model, introduced by Yu et al. (2025), utilizes Qwen2.5-7B-Base as its foundation and is trained on a mixture of datasets, including Math-PRM800K and Skywork-Reward-Preference-80K-v0.2. The training strategy combines Supervised Fine-Tuning (SFT) with Direct Preference Optimization (DPO), aiming to enhance judgment accuracy in decision-making tasks.

E Experiment Details

E.1 F1 and MCC Fairness Metrics

F1 and Matthews Correlation Coefficient Disparity. For each user $u \in \mathcal{U}_s$ and each prompt $x \in \mathcal{X}_u$, we derive pairwise preference labels as follows: The user-labeled binary classification set Y_s is constructed as:

$$Y_s = \bigcup_{u \in \mathcal{U}_s} \bigcup_{x \in \mathcal{X}_u} \{y_{u,x}^{(i,j)} \mid i < j \leq K_x\} \tag{6}$$

Similarly, the model-labeled binary classification set $\hat{Y}_s$ is constructed as:

$$\hat{Y}_{s,\theta} = \bigcup_{u \in \mathcal{U}_s} \bigcup_{x \in \mathcal{X}_u} \{\hat{y}_{\theta,x}^{(i,j)} \mid i < j \leq K_x\} \tag{7}$$

From Y_s and $\hat{Y}_{s,\theta}$, we compute classification metrics such as the binary or weighted F1 score $\mathrm{F1}_s = \mathrm{F1}(Y_s, \hat{Y}_{s,\theta})$ and the Matthews Correlation Coefficient $MCC_s = MCC(Y_s, \hat{Y}_{s,\theta})$ to quantify how well the model's predicted pairwise preferences align with the user-labeled ground truth, as well as the corresponding disparities across groups:

$$\Delta_{\text{F1}} = \max_{s_i, s_j \in \mathcal{S}, s_i \neq s_j} \left| \mathrm{F1}_{s_i} - \mathrm{F1}_{s_j} \right| \tag{8}$$

$$\Delta_{\text{MCC}} = \max_{s_i, s_j \in \mathcal{S}, s_i \neq s_j} \left| \mathrm{MCC}_{s_i} - \mathrm{MCC}_{s_j} \right| \tag{9}$$

Similarly, a higher disparity value for these metrics indicates a greater degree of unfairness.

E.2 Pearson's Chi-Squared Test

The Pearson's Chi-Squared test is a statistical method used to determine whether there is a significant association between categorical variables. It is commonly employed to test the independence between two variables or to assess the goodness-of-fit of an observed distribution compared to an expected distribution.

Given an observed frequency O_i and an expected frequency E_i for each category i, the chi-squared statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for category i,

[11] https://huggingface.co/R-I-S-E/RISE-Judge-Qwen2.5-7B

E_i is the expected frequency for category i, n is the number of categories.

The resulting chi-squared statistic follows a chi-squared distribution with $n-1$ degrees of freedom (for a goodness-of-fit test) or $(r-1)(c-1)$ degrees of freedom (for a test of independence), where r and c are the number of rows and columns, respectively. A significant result indicates that the observed frequencies deviate from the expected frequencies more than would be expected by random chance.

The hypothesis for the test can be stated as follows:

- H_0: The variables are independent (no association).
- H_1: The variables are not independent (there is an association).

E.3 Kruskal-Wallis Test

The Kruskal-Wallis test is a non-parametric statistical test used to determine whether there are statistically significant differences between the medians of three or more independent groups. It serves as a non-parametric alternative to the one-way ANOVA when the assumption of normality is violated.

The Kruskal-Wallis test statistic H is computed as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(R_i - \frac{N+1}{2} \right)^2$$

where N is the total number of observations across all groups, n_i is the Number of observations in group i, R_i is the Sum of ranks for group i, k is the Number of groups.

The null and alternative hypotheses are as follows:

- H_0: All groups have the same distribution (no difference in medians).
- H_1: At least one group has a different distribution.

The test statistic H approximately follows a chi-squared distribution with $k-1$ degrees of freedom under the null hypothesis. A significant p-value indicates that at least one group median significantly differs from the others.

E.4 Statistical Tests in Reward Model Evaluation

In the context of reward model evaluation, Pearson's chi-squared test is useful for testing if categorical preference agreements are independent, while the Kruskal-Wallis test is applied when comparing preference scores across multiple groups.

F Full Experiment Results

F.1 Full Preference Radar Plots across Demographic Groups

The full plots of stated preferences of different demographic groups can be found in Figure 6.

F.2 Kruskal-Wallis Test for Group Preferences

The Kruskal-Wallis p-value of each response quality axis across different demographic groups partitioned by a specific sensitive attribute can be found in Table 11.

F.3 Max-Min Score Difference for Each Attribute-Dimension Pair

We visualize the max-min score difference for each attribute-dimension pair in Figure 9.

F.4 Full Results of How Reward Models Capture Group Preferences

Extended from Section 4, we report the full results of how reward models capture group preferences in Table 14.

F.5 Full Results of How Reward Models Favor Group Perspectives

Extended from Section 5, we report the full result of how reward models favor group perspectives in Table 15.

Table 11: In Section 3, we compute the Kruskal-Wallis p-value across demographic groups on each response dimension. We observe that while all studied sensitive attributes exhibit some degree of preference variation, the most pronounced differences are associated with religion, geographic location, and ethnicity. On the response quality axes, the most pronounced differences are observed in values, creativity, safety, and diversity.

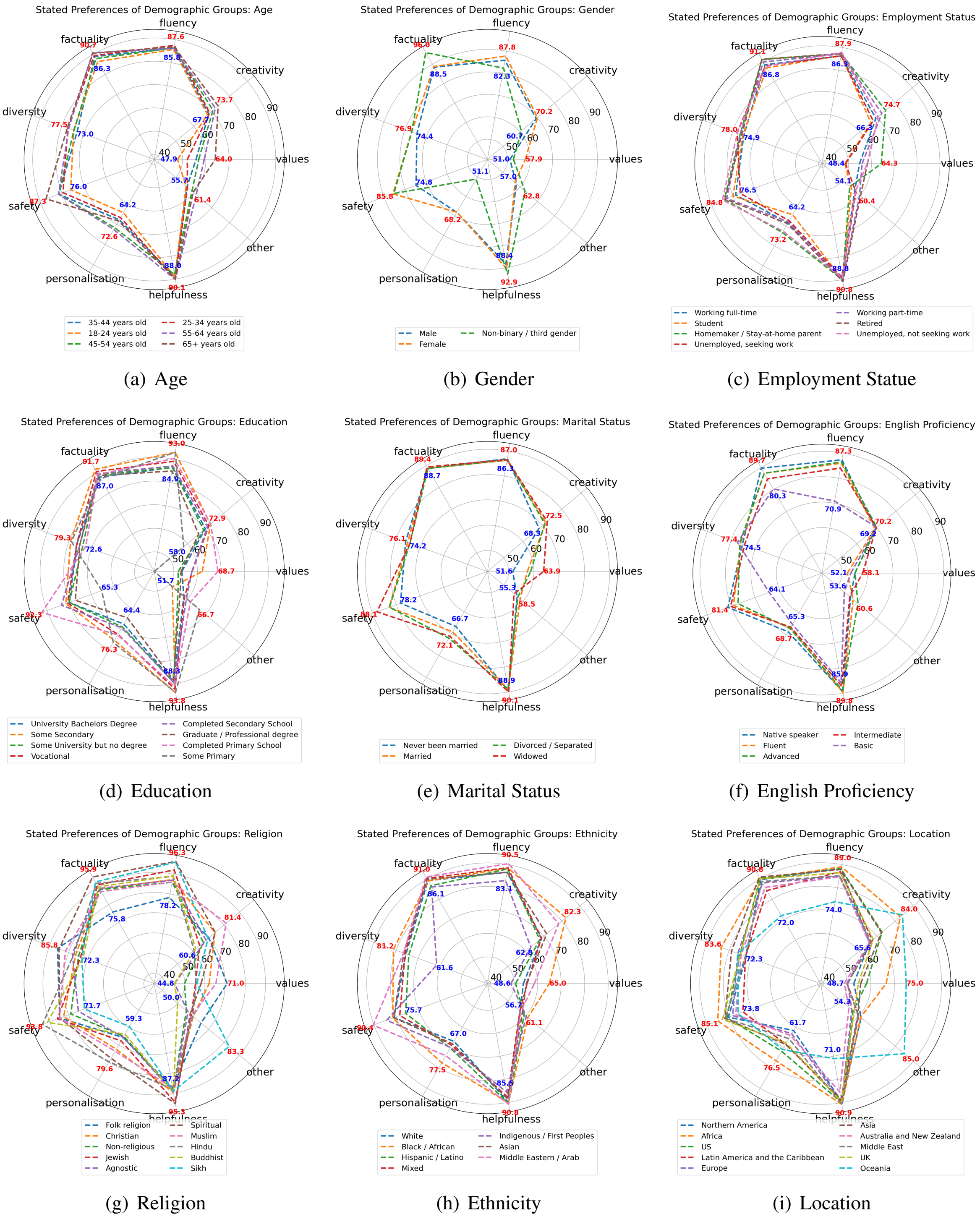
Sensitive Attribute	# Demographic Groups	Values	Creativity	Fluency	Factuality	Diversity	Safety	Personalisation	Helpfulness	Other
Age	6	4.94e-09	5.44e-02	9.89e-02	3.34e-03	1.01e-01	4.71e-04	3.00e-03	1.91e-01	3.26e-03
Gender	3	1.45e-05	1.42e-01	3.54e-03	1.47e-01	3.72e-02	3.49e-16	5.36e-02	5.50e-03	4.20e-01
Employment Status	7	2.75e-04	3.23e-02	9.63e-01	2.74e-01	7.20e-01	6.89e-03	1.81e-01	9.00e-01	6.27e-02
Education	8	1.67e-02	2.96e-01	9.53e-02	6.74e-01	6.90e-01	1.32e-01	4.35e-02	8.47e-02	6.00e-03
Marital Status	4	3.58e-05	1.33e-02	8.85e-01	9.55e-01	9.52e-01	1.28e-02	1.10e-01	6.90e-01	5.56e-01
English Proficiency	5	2.44e-01	9.91e-01	3.03e-01	2.03e-02	9.33e-01	2.57e-02	9.66e-01	6.07e-01	2.25e-01
Religion	10	1.64e-13	3.74e-08	2.62e-02	6.00e-01	4.10e-01	1.49e-03	2.06e-04	8.12e-02	4.53e-03
Ethnicity	7	4.12e-04	6.00e-11	3.97e-01	3.90e-01	2.57e-02	2.25e-01	2.26e-04	6.27e-01	1.89e-01
Location	10	7.03e-09	1.10e-13	6.26e-03	2.29e-02	3.31e-04	4.47e-04	7.31e-07	5.14e-02	4.53e-02

Table 12: Similar to Table 11 but rather than using stated preferences, we compute the Kruskal-Wallis p-value over the *performance preferences* (how well the chosen response performs) in the PRISM dataset.

Sensitive Attribute	# Demographic Groups	Values	Creativity	Fluency	Factuality	Diversity	Safety	Helpfulness
Age	6	5.14e-01	1.41e-01	4.03e-01	7.21e-01	8.90e-02	5.66e-01	4.41e-01
Gender	3	3.26e-02	1.41e-01	2.77e-03	3.03e-02	1.87e-01	1.40e-03	4.55e-03
Employment Status	7	3.20e-01	1.01e-01	2.03e-01	4.33e-01	2.77e-01	1.51e-01	1.95e-01
Education	8	5.85e-02	3.76e-02	2.15e-01	1.71e-01	3.72e-02	2.66e-01	5.33e-01
Marital Status	4	9.93e-02	1.87e-04	7.05e-01	3.52e-01	7.59e-03	6.38e-01	8.27e-01
English Proficiency	5	8.34e-01	2.07e-01	4.53e-01	5.57e-01	1.88e-01	5.05e-01	8.83e-01
Religion	10	7.87e-04	4.24e-07	8.68e-01	4.77e-02	1.19e-04	6.24e-01	4.83e-01
Ethnicity	7	2.95e-04	3.50e-05	1.15e-01	1.11e-02	1.20e-05	1.81e-02	9.58e-02
Location	9	5.51e-05	2.59e-09	1.00e-02	6.73e-05	5.80e-09	1.22e-01	4.66e-02

Table 13: Similar to Table 11 but rather than using stated preferences, we compute the Kruskal-Wallis p-value over the *choice preferences* (why use makes this choice over multiple responses) in the PRISM dataset.

Sensitive Attribute	# Demographic Groups	Values	Creativity	Fluency	Factuality	Diversity	Safety	Helpfulness
Age	6	2.69e-01	4.14e-01	5.96e-01	8.71e-01	3.61e-01	6.55e-01	6.42e-01
Gender	3	3.70e-03	2.33e-03	6.81e-04	6.03e-02	7.44e-03	1.66e-06	6.02e-05
Employment Status	7	3.65e-01	2.40e-01	4.90e-02	3.19e-01	6.31e-01	2.43e-01	7.69e-02
Education	8	3.45e-02	1.64e-02	2.46e-01	2.73e-01	1.84e-02	2.24e-03	2.82e-01
Marital Status	4	1.72e-02	1.14e-04	8.79e-01	4.89e-01	1.45e-02	4.67e-01	6.15e-01
English Proficiency	5	1.70e-01	1.22e-01	7.91e-01	4.20e-01	3.31e-01	1.32e-01	6.46e-01
Religion	10	3.34e-07	6.40e-09	2.05e-01	2.29e-03	7.98e-06	7.66e-04	3.32e-01
Ethnicity	7	1.24e-04	4.49e-05	1.30e-01	5.43e-02	1.43e-06	3.74e-03	2.38e-01
Location	9	8.71e-05	1.63e-07	4.99e-03	5.81e-03	8.62e-09	5.30e-05	2.70e-01


Stated Preferences of Demographic Groups: Age

(a) Age

Stated Preferences of Demographic Groups: Gender

(b) Gender

Stated Preferences of Demographic Groups: Employment Status

(c) Employment Statue

Stated Preferences of Demographic Groups: Education

(d) Education

Stated Preferences of Demographic Groups: Marital Status

(e) Marital Status

Stated Preferences of Demographic Groups: English Proficiency

(f) English Proficiency

Stated Preferences of Demographic Groups: Religion

(g) Religion

Stated Preferences of Demographic Groups: Ethnicity

(h) Ethnicity

Stated Preferences of Demographic Groups: Location

(i) Location

Figure 6: Full results of the stated preferences of different demographic groups across various LLM response quality dimensions (Kirk et al., 2024). The significance levels are validated in Table 11. From the results, certain demographic groups exhibit notably different preference patterns across various response quality dimensions. The Kruskal-Wallis p-value of each response quality axis across different demographic groups partitioned by a specific sensitive attribute can be found in Table 11.

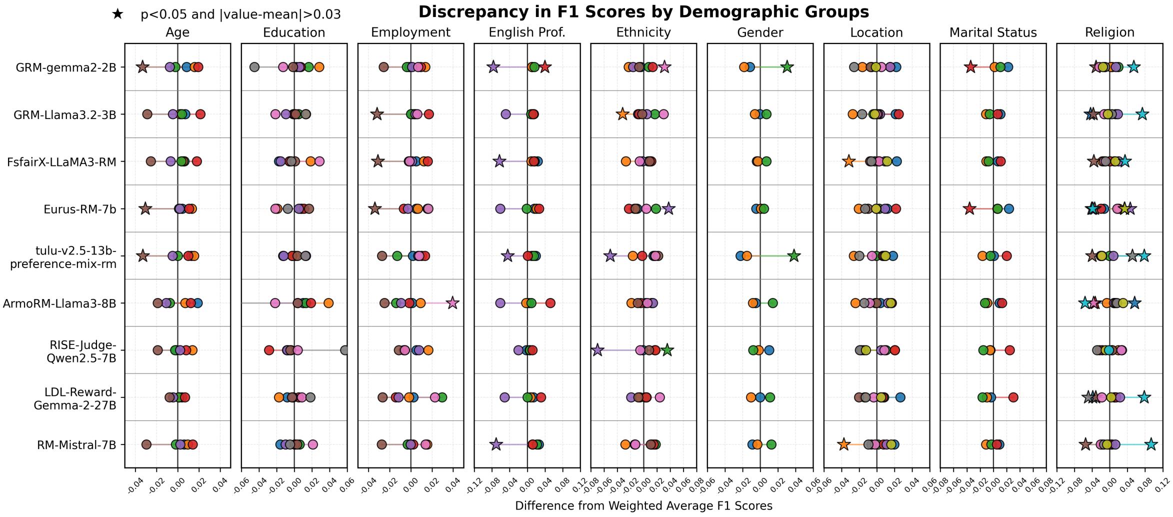


Figure 7: Similar to Figure 4 but using our developed F1 Fairness Metric (defined in Section E.1), we study the alignment between the reward models and different demographic groups. The values represent deviations from the average F1 score, with stars marking significantly unfair treatments. Groups within the same sensitive attribute are indicated by the same color.

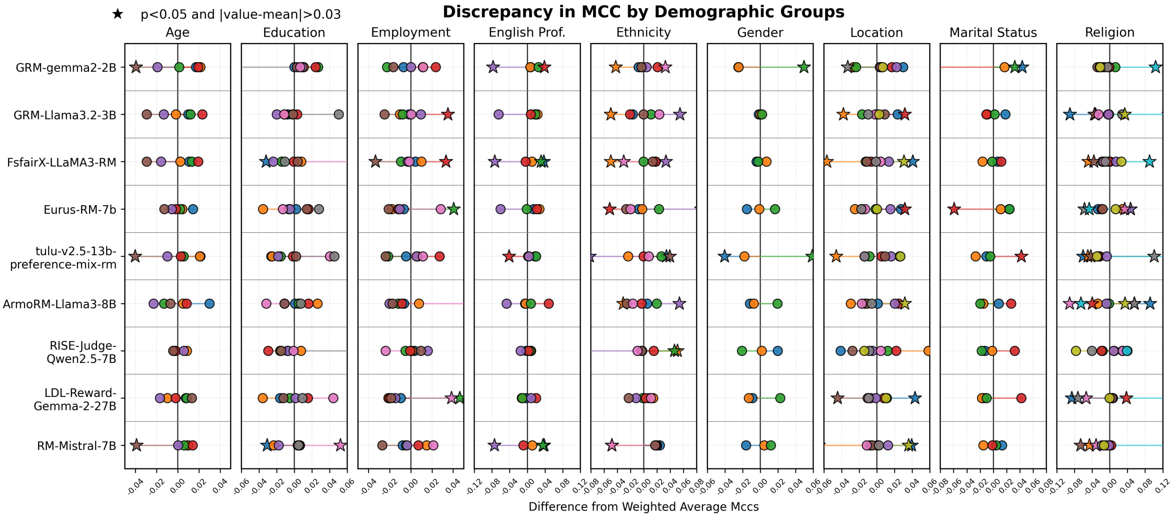


Figure 8: Similar to Figure 4 but using our developed Matthews Correlation Coefficient Fairness Metric (defined in Section E.1), we study the alignment between the reward models and different demographic groups. The values represent deviations from the average MCC score, with stars marking significantly unfair treatments. Groups within the same sensitive attribute are indicated by the same color.

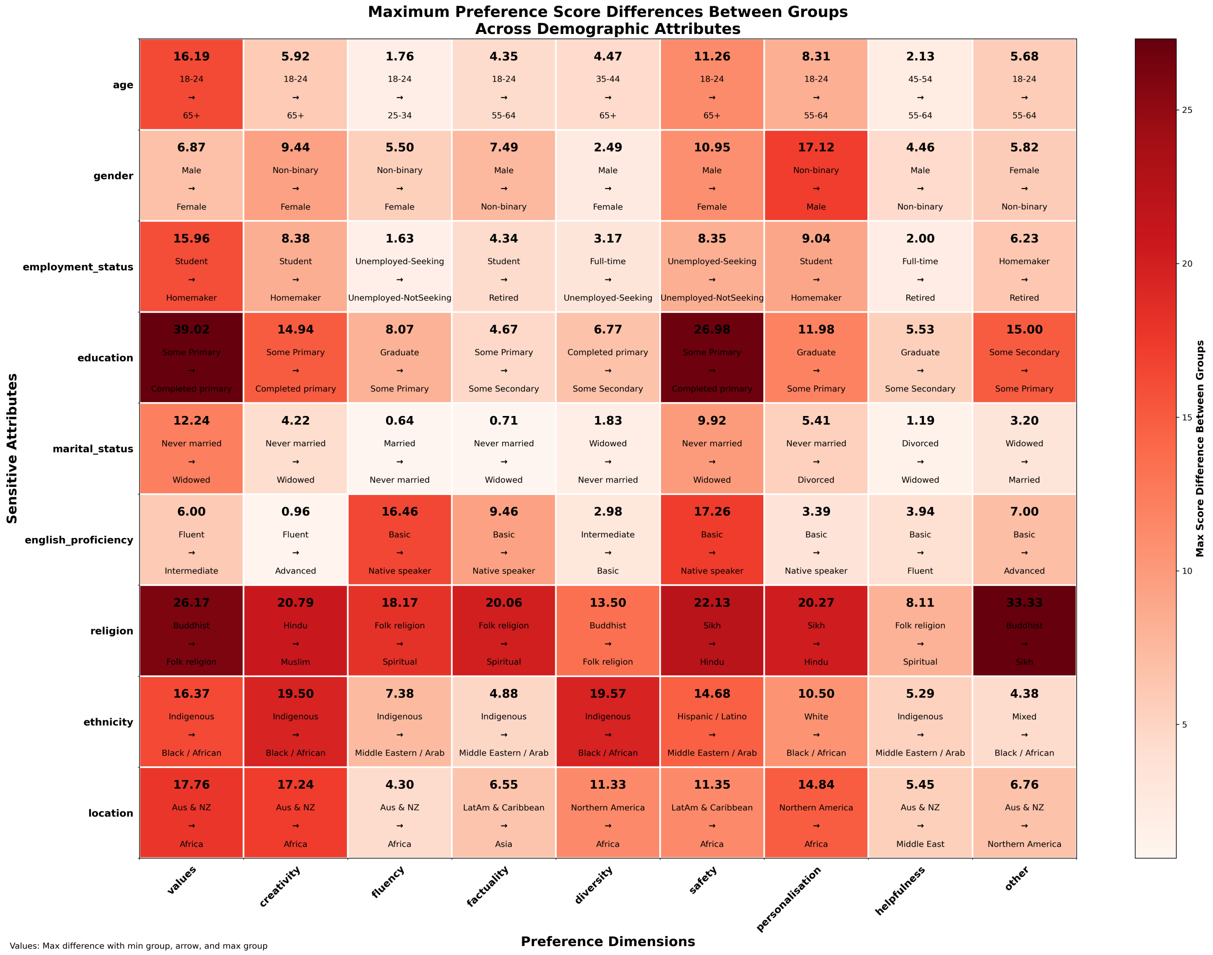


Figure 9: The max-min score difference for each attribute-dimension pair. The arrows highlight which max-min pair has the largest discrepancy. Darker red shades indicate more substantial divergence. We observe that certain groups tend to diverge more sharply in their stated preferences, and the most pronounced differences are associated with religion, geographic location, as well as ethnicity.

Table 14: Full result of SOTA reward models' fairness across demographic groups. Agreement disparity ($\Delta_{\text{agreement}}$) is defined in Section 4; binary and weighted F1 disparities (Δ_{F1} and $\Delta_{\text{F1-w}}$), as well as MCC disparity (Δ_{MCC}) are defined in Appendix E.1. Additionally, we perform Pearson's chi-squared test on the sets $\{Y_s, \hat{Y}_{s,\theta}\}_{s\in\mathcal{S}}$ to determine whether the reward model's alignment with user preferences significantly differs across groups. Red shades indicate significant p-values, emphasizing that different groups demonstrate varying levels of alignment with the reward model.

Reward Model	Size	Metric	Age	Gender	Employment Status	Education	Marital Status	English Proficiency	Religion	Ethnicity	Location
FsfairX-LLaMA3-RM-v0.1	8B	Agreement Disparity	4.4	0.8	3.6	4.3	2.0	9.0	10.0	5.8	7.3
		Binary F1 Disparity	4.3	1.1	4.7	4.6	2.2	8.8	7.1	3.8	5.4
		Weighted F1 Disparity	2.4	0.6	3.1	5.2	1.3	5.1	7.3	4.4	4.8
		MCC Disparity	4.9	1.1	6.7	10.3	2.8	11.1	13.9	8.3	9.7
		p-value	1.354×10^{-5}	1.682×10^{-1}	2.499×10^{-3}	2.923×10^{-3}	2.533×10^{-4}	1.296×10^{-5}	5.210×10^{-8}	9.782×10^{-12}	1.721×10^{-17}
ArmoRM-Llama3-8B-v0.1	8B	Agreement Disparity	2.0	2.3	5.3	3.2	2.9	8.4	8.4	4.5	5.6
		Binary F1 Disparity	3.7	2.3	6.4	10.5	2.6	11.4	11.3	3.2	4.1
		Weighted F1 Disparity	2.6	1.5	3.7	2.7	2.1	3.7	9.1	4.6	3.0
		MCC Disparity	5.3	3.1	7.7	5.8	4.7	9.5	18.3	8.5	6.1
		p-value	3.000×10^{-2}	3.606×10^{-1}	7.082×10^{-4}	9.205×10^{-1}	1.664×10^{-3}	8.767×10^{-2}	7.702×10^{-6}	1.067×10^{-6}	3.742×10^{-10}
GRM-Llama3.2-3B-rewardmodel-ft	3B	Agreement Disparity	4.6	1.1	3.5	5.7	2.3	7.4	12.7	6.0	6.2
		Binary F1 Disparity	5.1	1.3	5.0	3.5	2.1	6.4	11.7	6.2	5.3
		Weighted F1 Disparity	2.5	0.1	2.7	3.7	1.7	4.3	10.5	5.6	3.3
		MCC Disparity	5.2	0.3	6.0	7.1	2.8	8.8	22.4	10.4	6.9
		p-value	4.150×10^{-6}	7.225×10^{-1}	1.202×10^{-2}	7.671×10^{-1}	6.434×10^{-6}	8.750×10^{-2}	1.950×10^{-6}	6.458×10^{-8}	1.171×10^{-10}
RM-Mistral-7B	7B	Agreement Disparity	5.1	1.9	4.5	3.5	2.2	11.1	14.9	6.2	8.1
		Binary F1 Disparity	4.4	2.1	4.3	3.6	2.0	9.7	14.9	4.6	5.7
		Weighted F1 Disparity	2.6	1.3	2.3	4.2	1.4	4.9	11.9	3.7	5.2
		MCC Disparity	5.3	2.8	4.8	8.3	2.9	11.2	25.1	7.2	10.5
		p-value	4.276×10^{-7}	1.332×10^{-2}	3.025×10^{-4}	1.626×10^{-3}	9.218×10^{-6}	5.345×10^{-6}	8.430×10^{-10}	2.003×10^{-13}	3.503×10^{-20}
Eurus-RM-7b	7B	Agreement Disparity	3.1	2.7	4.1	5.5	5.1	6.6	9.9	4.2	5.1
		Binary F1 Disparity	4.4	0.9	5.1	3.8	5.9	8.8	8.7	6.0	4.2
		Weighted F1 Disparity	1.1	1.7	3.1	3.5	4.2	3.8	5.1	6.4	2.7
		MCC Disparity	2.7	3.2	6.2	6.3	8.3	8.8	10.5	13.8	5.7
		p-value	2.586×10^{-2}	1.152×10^{-1}	2.501×10^{-2}	4.818×10^{-1}	1.682×10^{-4}	3.871×10^{-1}	1.035×10^{-4}	4.813×10^{-3}	6.623×10^{-7}
RISE-Judge-Qwen2.5-7B	7B	Agreement Disparity	2.8	1.0	1.3	4.4	3.3	5.4	10.1	8.4	2.7
		Binary F1 Disparity	3.3	1.8	2.8	8.6	4.0	3.3	5.7	10.6	3.9
		Weighted F1 Disparity	0.5	2.0	2.1	4.1	2.3	1.2	6.5	7.1	5.1
		MCC Disparity	1.2	4.1	4.0	9.9	5.1	2.3	11.6	14.7	10.0
		p-value	1.159×10^{-2}	4.672×10^{-2}	7.065×10^{-1}	6.050×10^{-1}	2.067×10^{-1}	5.166×10^{-1}	2.168×10^{-1}	8.415×10^{-3}	2.626×10^{-1}
tulu-v2.5-13b-preference-mix-rm	13B	Agreement Disparity	5.4	5.6	3.2	6.0	3.8	6.7	14.7	6.9	6.8
		Binary F1 Disparity	4.8	6.0	4.0	2.6	3.7	6.3	11.9	7.3	4.4
		Weighted F1 Disparity	3.0	4.8	2.4	3.8	3.5	3.1	11.5	6.3	3.7
		MCC Disparity	6.2	9.9	5.1	7.1	6.9	6.1	22.7	12.1	7.3
		p-value	3.522×10^{-7}	4.038×10^{-5}	2.412×10^{-3}	6.381×10^{-2}	6.512×10^{-4}	1.701×10^{-4}	1.848×10^{-6}	6.726×10^{-12}	4.124×10^{-14}
LDL-Reward-Gemma-2-27B-v0.1	27B	Agreement Disparity	1.0	2.7	4.7	4.4	3.1	4.2	16.1	3.5	5.3
		Binary F1 Disparity	1.4	2.1	5.6	3.6	4.7	8.2	12.7	4.3	4.7
		Weighted F1 Disparity	1.7	1.8	3.3	4.2	2.9	2.0	11.9	2.0	4.4
		MCC Disparity	3.0	3.6	6.7	8.0	6.0	3.3	24.3	3.7	8.8
		p-value	7.561×10^{-1}	2.560×10^{-1}	3.570×10^{-3}	1.527×10^{-1}	7.324×10^{-2}	7.892×10^{-2}	2.351×10^{-4}	2.317×10^{-3}	9.076×10^{-6}
GRM-gemma2-2B-rewardmodel-ft	2B	Agreement Disparity	5.0	4.6	3.2	2.5	7.3	10.4	7.8	6.6	5.8
		Binary F1 Disparity	5.2	4.9	3.9	7.4	5.7	11.7	8.6	5.4	4.8
		Weighted F1 Disparity	2.9	3.7	2.5	5.3	7.0	5.0	6.0	3.8	3.0
		MCC Disparity	6.1	7.4	4.6	10.7	13.6	11.6	13.1	7.5	6.3
		p-value	5.462×10^{-8}	2.004×10^{-2}	5.112×10^{-3}	3.313×10^{-1}	2.034×10^{-8}	2.189×10^{-2}	3.024×10^{-4}	8.391×10^{-8}	3.823×10^{-12}

Table 15: Mean reward scores across demographic groups by sensitive attribute, with p-values indicating statistical significance of group-wise differences.

Sensitive Attribute	Demographic Group	Mean Reward	p-value
Age	18–24 years old	-5.4316	0.0492
	25–34 years old	-5.3692	
	35–44 years old	-5.3939	
	45–54 years old	-5.5608	
	55–64 years old	-5.6355	
	65+ years old	-5.5849	
Gender	Male	-5.5670	0.0303
	Female	-5.4682	
	Non-binary / third gender	-6.0213	
Employment	Working full-time	-5.2809	1.42e-13
	Student	-5.9489	
	Homemaker / Stay-at-home	-6.5225	
	Unemployed, seeking work	-5.6273	
	Working part-time	-5.7092	
	Retired	-5.7821	
	Unemployed, not seeking work	-5.3682	
Education	University Bachelor's Degree	-5.0576	0.0345
	Some Secondary	-5.2874	
	Some University (no degree)	-5.0488	
	Vocational	-5.3245	
	Completed Secondary School	-4.9181	
	Graduate / Professional	-5.1166	
	Completed Primary School	-5.0972	
	Some Primary	-5.3016	
Marital Status	Never been married	-5.4125	0.2477
	Married	-5.4263	
	Divorced / Separated	-5.7487	
	Widowed	-5.8701	
English Proficiency	Native speaker	-5.0814	0.0034
	Fluent	-4.9285	
	Advanced	-4.8378	
	Intermediate	-5.1255	
	Basic	-5.3757	
Religion	Folk religion	-7.2228	9.48e-13
	Christian	-6.9378	
	Non-religious	-6.3325	
	Jewish	-6.7394	
	Agnostic	-6.3602	
	Spiritual	-6.7769	
	Muslim	-7.0822	
	Hindu	-6.9106	
	Buddhist	-7.2275	
	Sikh	-7.0338	
Ethnicity	White	-5.4721	3.13e-05
	Black / African	-5.7886	
	Hispanic / Latino	-5.8407	
	Mixed	-5.8511	
	Indigenous / First Peoples	-6.3272	
	Asian	-5.8778	
	Middle Eastern / Arab	-5.9991	
Location	Northern America	-5.3430	4.59e-13
	Africa	-5.9212	
	US	-5.2954	
	Latin America / Caribbean	-6.0999	
	Europe	-5.4743	
	Asia	-5.4445	
	Australia / New Zealand	-5.8822	
	Middle East	-6.0614	
	UK	-5.6950	
	Oceania	-5.8178	