

Towards Reward Fairness in RLHF: From a Resource Allocation Perspective

Sheng Ouyang^{1,2,3,4*}, Yulan Hu^{4†}, Ge Chen^{4,5*}, Qingyang Li⁴, Fuzheng Zhang⁴, Yong Liu^{1,2,3,4†}

¹Gaoling School of Artificial Intelligence, Renmin University of China

²Beijing Key Laboratory of Research on Large Models and Intelligent Governance

³Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE

⁴Kuaishou Technology

⁵University of Chinese Academy of Sciences

{ouyangsheng, liuyonggsai}@ruc.edu.cn, {huyulan, liqingyang, zhangfuzheng}@kuaishou.com, chenge221@mails.ucas.ac.cn

Abstract

Rewards serve as proxies for human preferences and play a crucial role in Reinforcement Learning from Human Feedback (RLHF). However, if these rewards are inherently imperfect, exhibiting various biases, they can adversely affect the alignment of large language models (LLMs). In this paper, we collectively define the various biases present in rewards as the problem of reward unfairness. We propose a bias-agnostic method to address the issue of reward fairness from a resource allocation perspective, without specifically designing for each type of bias, yet effectively mitigating them. Specifically, we model preference learning as a resource allocation problem, treating rewards as resources to be allocated while considering the trade-off between utility and fairness in their distribution. We propose two methods, Fairness Regularization and Fairness Coefficient, to achieve fairness in rewards. We apply our methods in both verification and reinforcement learning scenarios to obtain a fairness reward model and a policy model, respectively. Experiments conducted in these scenarios demonstrate that our approach aligns LLMs with human preferences in a more fair manner. Our data and code are available at <https://github.com/shoyua/Towards-Reward-Fairness>.

1 Introduction

RLHF (Ouyang et al., 2022; Kaufmann et al., 2023; Dong et al., 2024) has significantly advanced the alignment of LLM outputs with human preferences, ensuring that the responses are helpful, harmless, and honest (Bai et al., 2022; Huang et al., 2024). The reward model (RM) (Stiennon et al., 2020; Ouyang et al., 2022; Yan et al., 2024) plays a crucial role in this process by providing a quantitative metric that measures the degree to which the model

* Work done during an internship at Kuaishou Technology.

† Corresponding authors.

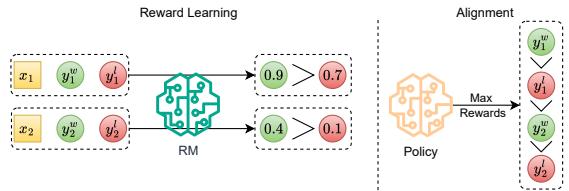


Figure 1: Rewards unfair problem in RLHF.

outputs align with human preferences. This metric guides the LLM in producing outputs that are more consistent with human preference.

One of the key reasons for the success of RLHF is the assumption that the reward model can accurately represent and measure actual preferences (Kim and Seo, 2024). However, if the reward model itself is biased (Son et al., 2024; Hayes et al., 2024; Reber et al., 2024), it can lead to policy models exhibiting behaviors that do not align with human preferences (Gao et al., 2023; Dubois et al., 2024b). Consider Figure 1, where we denote a preference pair data as (x, y^w, y^l) , representing a query, a preferred and dispreferred response, respectively. When training a reward model following the Bradley-Terry model (Bradley and Terry, 1952) with such preference data, the output appears reasonable because for each x , the reward of y^w is greater than that of y^l . However, using this reward model to guide the training of a policy model can be problematic. The policy model aims to maximize rewards, and in this scenario, the reward for y_1^l is greater than the reward for y_2^w , even though y_1^l is dispreferred and y_2^w is preferred. From a general perspective, if (x_1, y_1^w, y_1^l) and (x_2, y_2^w, y_2^l) come from different data types, such bias in their reward distribution can steer the model to favor one type of data over another. We define this issue as “**reward unfairness**”.

We interpret various reward biases from the perspective of reward unfairness, including length bias (Shen et al., 2023; Park et al., 2024), category

bias (Padmakumar et al., 2024), and social bias (Li et al., 2023). For instance, when the rewards distribution varies significantly across data of different lengths or categories, resulting in reward unfairness, it manifests as length bias or category bias respectively. Existing work (Park et al., 2024; Chen et al., 2024) has addressed these biases by proposing targeted methods to mitigate them. Park et al. (2024); Chen et al. (2024); Yang et al. (2024); Padmakumar et al. (2024) have employed techniques such as length regularization to mitigate the effects of length bias. These methods adjust the reward distribution to prevent models from favoring longer responses, thereby ensuring more fair outputs. On the other hand, category bias has not been as widely acknowledged. However, some studies have implicitly addressed this issue. For instance, the work on learning diverse preferences (Yang et al., 2024; Padmakumar et al., 2024) and model ensemble (Ramé et al., 2024b,a) indirectly reduce the impact of category bias. These work promote a more varied and representative set of outputs, which helps minimize the skewness introduced by category-specific biases.

However, these works are specifically designed to address particular biases and lack the ability to transfer solutions across different types of biases. In this paper, we propose a unified perspective that considers these biases as manifestations of a broader issue: reward unfairness. To address this comprehensively, we introduce the reward fairness framework. Firstly, we model preference learning as a resource allocation problem (Katoh and Ibaraki, 1998). In this framework, we define the rewards in preference learning as the resources to be allocated. The extent to which these rewards reflect human preferences is defined as utility, while the consistent distribution of rewards across the data is defined as fairness. We employ a unified fairness function to measure the fairness of the rewards distribution. This approach seeks to achieve a trade-off between fairness and utility. We propose two methods to obtain fairness rewards: Fairness Regularization and Fairness Coefficient. We then apply these methods in two scenarios: Fairness Rewards for Verification and Fairness Rewards for Reinforcement Learning (RL). We conclude our contributions as following:

Unified Perspective from Reward Unfairness

We introduce a novel perspective that frames various biases as specific instances of the broader prob-

lem of reward unfairness. This unified view fosters a more comprehensive understanding and approach to addressing these biases.

Reward Fairness Framework We propose the reward fairness framework from a resource allocation perspective to systematically address reward unfairness, aiming to balance fairness and utility in reward distribution.

Application to Verification and RL We apply our proposed methods in two scenarios: (a) Fairness Rewards for Verification, which focuses on training a fairness RM, and (b) Fairness Rewards for RL, which aims to train a policy model that implicitly incorporates fair rewards. Our fairness rewards methods can be seamlessly integrated with existing RM and RL methods.

2 Related Work

RLHF has become the standard approach for aligning LLMs with human preferences. RLHF can be decomposed into two main components: Reward Learning and RL Finetune.

2.1 Reward Learning

The reward model is a crucial component of RLHF, providing a quantitative metric to guide alignment with human preferences. Reward models typically follow the Bradley-Terry model (Bradley and Terry, 1952), but there are also approaches based on regression paradigms (Wang et al., 2024a,b) and the “LLM as a judge” approach (Zhang et al., 2024; Zheng et al., 2023). However, Hou et al. (2021); Kim and Seo (2024); Reber et al. (2024) have identified that reward models are imperfect proxies for human preferences, exhibiting various issues such as length bias (Shen et al., 2023) and reward hacking (Skalse et al., 2022). Shen et al. (2023); Chen et al. (2024) have found that the results of reward models are influenced by the length of the input, and they have attempted to decouple this relationship during training to mitigate its effects. Fast RL (Li et al., 2024) is closest to our method, however Fast RL is an ensemble method that considers fairness between different reward functions.

2.2 RL Finetune

RL Finetuning (Dong et al., 2024) generally involves using reinforcement learning techniques, guided by the reward model, to train the policy model. Algorithms such as proximal policy optimization (PPO) (Schulman et al., 2017) and group

relative policy optimization (GRPO) (Shao et al., 2024) are commonly used. There is also a category of work that omits the reward model and directly learns preference, such as direct preference optimization (DPO) (Rafailov et al., 2024), Kahneman-Tversky optimization (KTO) (Ethayarajh et al., 2024), and SimPO (Meng et al., 2024). These methods are more efficient and stable compared to PPO-based approaches. Although they do not involve the reward model in training, they implicitly fit rewards to align with human preferences. Lu et al. (2024); Liu et al. (2024); Dubois et al. (2024a) have observed that aligned models tend to generate longer responses, which introduces a length bias. To mitigate this issue, they have proposed methods such as length regularization (Park et al., 2024).

3 Preliminaries

Reward Model In RLHF, RM acts as a proxy for human preferences to rate the quality of the model output. Generally, the RM follows the Bradley-Terry Model (Bradley and Terry, 1952) and can be formulated as:

$$p(y_w \succ y_l | x) = \frac{\exp(r_\phi(x, y_w))}{\exp(r_\phi(x, y_w)) + \exp(r_\phi(x, y_l))}, \quad (1)$$

where $(x, y_w, y_l) \sim \mathcal{D}$ represent a prompt, a preferred response and an dispreferred response from the preference dataset \mathcal{D} , respectively. $r_\phi(x, y)$ denotes a reward function with the parameters ϕ , and this is subsequently denoted as $r_\phi(y)$ for simplicity. We can train a RM r_ϕ following the log-likelihood maximization as:

$$\max_{r_\phi} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(y_w) - r_\phi(y_l))], \quad (2)$$

where σ is the sigmoid function.

RL Finetune During the RL phase (Jaques et al., 2017), the learned RM is used to provide feedback to the policy model π_θ with the parameters θ . The optimization is formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(y)] - \beta D_{\text{KL}}[\pi_\theta(y|x) \| \pi_{\text{ref}}(y|x)], \quad (3)$$

where β is a hyperparameter controlling the KL penalty and π_{ref} is the reference model.

DPO DPO (Rafailov et al., 2024) is a method used to directly optimize a policy based on preference data. The objective of DPO is to align the policy π_θ with human preferences by maximizing the likelihood of preferred outcomes.

$$\max_{\pi_\theta} \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right], \quad (4)$$

where β is a scaling factor.

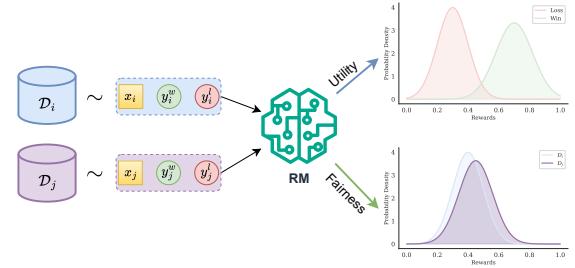


Figure 2: Objective of Fairness Rewards. \mathcal{D}_i and \mathcal{D}_j represent different data. Fairness rewards aim to obtain rewards that consider the trade-off between utility and fairness. Utility refers to the ability of the rewards to distinguish between preferred and dispreferred responses, as illustrated in the top-right figure. Fairness refers to the consistent distribution of rewards across different data, as depicted in the bottom-right figure.

4 Fairness Rewards Allocation

In this section, we aim to obtain the fairness rewards, as shown in Figure 2. We model preference learning as a resource allocation problem that maximizes utility while ensuring the fairness of rewards.

Resource Allocation Resource allocation (Katoh and Ibaraki, 1998) involves distributing resources R among entities i to optimize overall utility U . The trade-off between fairness and utility is a key consideration. Utility maximization is given by:

$$\max U(\mathbf{a}) \quad \text{s.t.} \quad \sum_i a_i \leq R, \quad (5)$$

where $\mathbf{a} = [a_1, a_2, \dots, a_n]$ is an allocation vector and a_i denotes the resources allocated to the i -th entity. Fairness can be incorporated by adding a fairness constraint $F(\mathbf{a})$:

$$\begin{aligned} \max & \quad U(\mathbf{a}) \\ \text{s.t.} & \quad \sum_i a_i \leq R, \\ & \quad F(\mathbf{a}) \geq \eta \end{aligned} \quad (6)$$

where η represents the desired level of fairness. Balancing these objectives requires careful consideration of both fairness and utility to achieve an optimal allocation strategy.

Fairness Rewards In RLHF, rewards are quantified representations of human preferences, reflecting the degree to which model outputs align with human preferences. We model preference learning as a resource allocation problem, where rewards are considered as resources to be allocated. According to Eq (5), our objective is to maximize the utility of reward allocation. We use $U(\mathbf{a})$ to measure the extent to which the reward allocation vector \mathbf{a} aligns with human preferences, with higher values indicating greater utility. Concurrently, it is imperative to ensure fairness in the distribution of rewards. We use $F(\mathbf{a})$ to measure the fairness of the reward allocation vector, with larger values indicating greater fairness. We expect $F(\mathbf{a})$ to satisfy following properties:

1. **Continuity** The fairness measure $F(\mathbf{a})$ is continuous on \mathbb{R}_+^n for all integers $n \geq 1$.

This property ensures that small changes in resource allocation result in only minor changes to the fairness measure, thereby guaranteeing the stability and consistency of the fairness measure.

2. **Homogeneity** The fairness measure $F(\mathbf{a})$ is a homogeneous function of degree 0:

$$F(\mathbf{a}) = F(t \cdot \mathbf{a}), \forall t > 0.$$

This property indicates that the fairness measure is independent of the scale of resource allocation.

3. **Monotonicity** For $n = 2$ entities, the fairness measure $F(\theta, 1 - \theta)$ is monotonically increasing as the absolute difference between the two elements (i.e., $|1 - 2\theta|$) shrinks to zero.

This property states that the fairness measure increases as the resource allocation between two entities becomes more equal.

There is a unified fairness metric proposed by Lan et al. (2010) that satisfies three key properties:

$$f_\tau(\mathbf{a}) = \text{sign}(1 - \tau) \cdot \left[\sum_{i=1}^n \left(\frac{a_i}{\sum_j a_j} \right)^{1-\tau} \right]^{\frac{1}{\tau}}, \quad (7)$$

where $\tau \in \mathbb{R}$ is a constant, which allows for the derivation of different fairness functions based on its value. For instance, when $\tau = -1$, $f_{\tau=-1}(\mathbf{a}) = \frac{(\sum_i a_i)^2}{\sum_i a_i^2} = n \cdot J(\mathbf{a})$ results in Jain's index $J(\mathbf{a})$ (Jain et al., 1984), which is a famous metric for measuring fairness in the resource allocation.

According to Eq (6), we consider the trade-off between utility and fairness. We employ Eq (7) to measure the fairness of reward allocation. Since rewards can be viewed as an infinite resource, and given the property of Homogeneity in fairness metrics, the fairness measure is independent of the unit of measurement or the size of the resource allocation. We can eliminate the constraint on the total amount of resources from Eq (6). Consequently, we propose the following two methods to transform Eq (6) into an unconstrained optimization problem.

- **Fairness Regularization:** We add the two measures together,

$$\max U(\mathbf{a}) + \alpha F(\mathbf{a}), \quad (8)$$

where α is a hyperparameter that controls the impact of the fairness regularization.

- **Fairness Coefficient:** We multiply the two measures,

$$\max U(\mathbf{a}) \cdot F(\mathbf{a})^\gamma, \quad (9)$$

where γ is a hyperparameter that controls the impact of the fairness coefficient.

By incorporating these methods, we aim to achieve a trade-off between fairness and utility in the rewards allocation process, ensuring that the rewards not only reflect human preferences accurately but also do so in a fair manner.

Clarification of Fairness Finally, we clarify that in the context of LLMs, “fairness” usually relates to “social bias” (Li et al., 2023; Gallegos et al., 2024). However, in this paper, we reformulate preference learning from a resource allocation perspective, treating rewards as allocated resources. Here, “fairness” refers to the fairness of reward allocation, drawing from resource allocation literature (Kumar and Kleinberg, 2000; Lan et al., 2010), which differs from social bias. As stated in Section 1, we interpret various reward biases through the lens of reward unfairness, addressing them uniformly. For example, in length bias, “entity” refers to data of varying lengths; for category

bias, it refers to different data categories, such as “helpful” and “harmless”. In social bias cases, like gender bias, “entity” denotes different genders.

5 Reward-Fairness RLHF

In this section, we discuss the application of fairness rewards in two scenarios:

- Fairness Rewards for Verification (§5.1): We introduce how to train a reward-fairness reward model, which serves as a fair verification.
- Fairness Rewards for RL (§5.2): We detail the training of a reward-fairness policy model, which aims to generate outputs that are fairer.

5.1 Fairness Rewards for Verification

The objective of the reward model is to act as a proxy for human preferences. Typically, it takes a pair of prompt and response (x, y) as input and outputs a scalar score to verify the quality of the pair. To develop a reward-fairness reward model, we need to define the utility function U and the fairness function F as per Eq (8) and (9).

For the Bradley-Terry reward model, which uses Eq (2) as its training objective, the goal is to allocate a higher reward to the preferred response y_w compared to the dispreferred response y_l . Therefore, we define the elements of the allocation vector a_i as $a_i = r_\phi(y_w) - r_\phi(y_l)$. With the allocation vector \mathbf{a} defined, we can directly take Eq (7) as the fairness function, i.e., $F(\mathbf{a}) = f_\tau(\mathbf{a})$. The utility function for the reward model is then defined as:

$$U(\mathbf{a}) = \mathbb{E}_{a_i \in \mathbf{a}} [\log \sigma(a_i)]. \quad (10)$$

We define two types of reward models incorporating fairness: Reward Model with Fairness Regularization (FR RM) and Reward Model with Fairness Coefficient (FC RM). Their training objectives are as follows:

FR RM The training objective combines the utility and fairness measures additively:

$$\mathcal{L}_{\text{FR RM}} = -\mathbb{E}_{a_i \in \mathbf{a}} [\log \sigma(a_i)] - \alpha F(\mathbf{a}). \quad (11)$$

FC RM The training objective combines the utility and fairness measures multiplicatively:

$$\mathcal{L}_{\text{FC RM}} = -\mathbb{E}_{a_i \in \mathbf{a}} [\log \sigma(a_i)] \cdot F(\mathbf{a})^\gamma. \quad (12)$$

5.2 Fairness Rewards For RL

Although the training of DPO does not explicitly involve a reward model, it implicitly fits a reward model (Rafailov et al., 2024). We can interpret the term $\beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ as an implicit reward. Similar to the reward model, we define the elements of the allocation vector a_i as:

$$a_i = \beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)}. \quad (13)$$

Consequently, we can derive the same utility function as for the reward model, as shown in Eq (10). The utility function of the DPO has the same form as that of the reward model, except that the meaning of the allocation vector is slightly different, where $a_i \in \mathbf{a}$ represents the difference in implicit rewards between the preferred and less preferred responses. With the allocation vector, we can directly use Eq (7) as a fairness function.

We define two types of DPO models incorporating fairness: DPO with Fairness Regularization (FR DPO) and DPO with Fairness Coefficient (FC DPO). The training objectives of $\mathcal{L}_{\text{FR DPO}}$ and $\mathcal{L}_{\text{FC DPO}}$ can be converted to the same form as Eq (11) and (12) respectively.

By incorporating these fairness measures into the DPO framework, we aim to ensure that the model not only aligns with human preferences but also allocate implicit rewards in a fair manner.

6 Experiments

In this section, we empirically investigate the following two research questions \mathcal{RQ} :

- $\mathcal{RQ}1$: how effective is our Fairness Rewards approach in both verification and RL scenarios?
- $\mathcal{RQ}2$: how does the choice of Fairness Function impact performance?

Datasets & Baselines In the verification scenario, we conduct experiments on two benchmarks: Reward Bench (Lambert et al., 2024) and HH-RLHF (Bai et al., 2022). Our RMs are trained using the training set from HH-RLHF, making HH-RLHF an in-distribution (ID) benchmark, while Reward Bench serves as an out-of-distribution (OOD) benchmark. We report the accuracy on both benchmarks. For the RL scenario, we evaluate our methods on the AlpacaEval2 (Dubois et al., 2024a) and

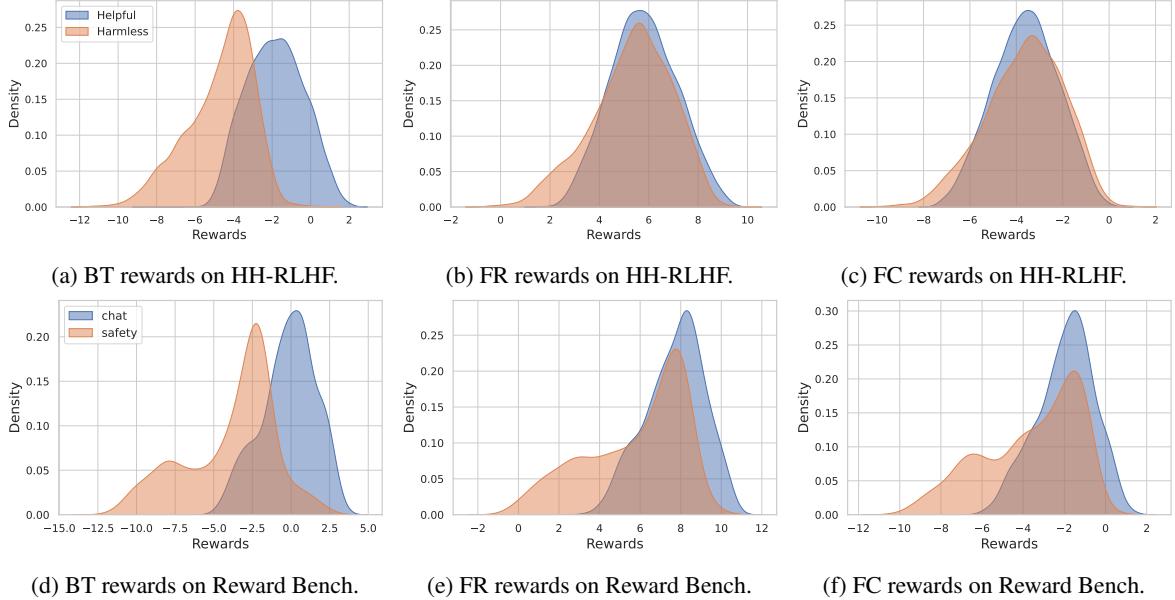


Figure 3: Rewards across ID and OOD data.

MT-Bench (Zheng et al., 2023) benchmarks. We provide results for AlpacaEval2 in terms of Length-controlled Win Rate (LC WR) and Win Rate (WR), and for MT-Bench, we report the overall score. For the policy model, we utilize the UltraFeedback Binarized and SHP datasets for training. We train the policy model using different methods such as DPO (Rafailov et al., 2024), KTO (Ethayarajh et al., 2024) and R-DPO (Park et al., 2024) with HALOs.

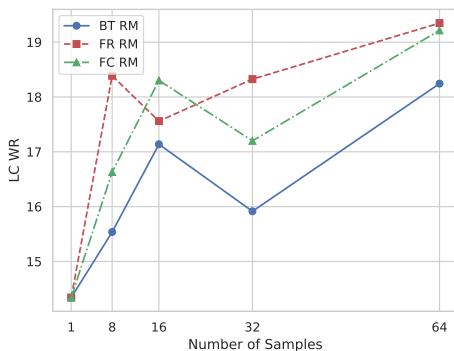


Figure 4: Performance of LLaMA3-SFT on AlpacaEval2 using different verification strategies.

Implementation Details For the reward model, we train on the HH-RLHF training set for one epoch with a learning rate of 2×10^{-6} . For the policy model, we utilize the UltraFeedback Binarized and SHP datasets, training for one epoch with a learning rate of 5×10^{-6} . During sampling with the policy model, the temperature coefficient is set to 1. All experiments are performed on an 8 ×

H800 machine. Both the reward models and the policy models are trained using LLaMA3-SFT (a base model developed by Dong et al. (2024)) and Qwen2.5-SFT (a base model we trained following Dong et al. (2024)). Further experimental details can be found in Appendix B.

6.1 Main Results (\mathcal{RQ}_1)

6.1.1 Fairness Verification

Figure 3 illustrates the distribution of rewards from different RMs on ID data (HH-RLHF) and OOD data (Reward Bench). The first row of figures shows the rewards distribution on the ID dataset HH-RLHF. It is evident that the Bradley-Terry (BT) RM exhibits a significant disparity in the distribution of rewards between Helpful and Harmless data, indicating an unfair allocation of rewards. In contrast, Reward Model with Fairness Regularization (FR RM) and Reward Model with Fairness Coefficient (FC RM) demonstrate a more consistent rewards distribution across Helpful and Harmless data, indicating that those two RMs are fairer. The figures in the second row show the distribution of rewards on the OOD data, and we can draw the same conclusions as for the ID data. Table 1 presents the performance of the three RMs on the Reward Bench and HH-RLHF. The results show no significant performance difference between FR RM, FC RM, and BT RM, suggesting that Fair RMs achieve a good trade-off between fairness and utility without sacrificing model performance. Additionally, we provide the distribution of rewards on

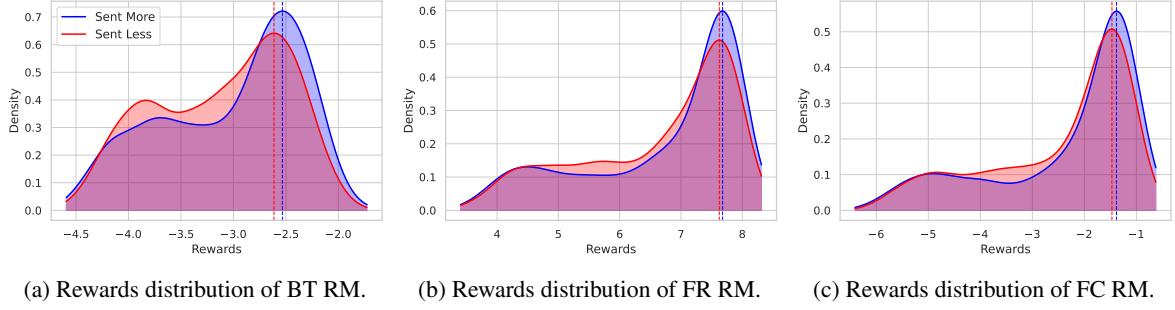


Figure 5: Rewards distribution on CrowS-Pairs.

Table 1: Performance of different verification strategies on two reward model benchmarks.

Verifiers	Reward Bench					HH-RLHF		
	Chat	Chathard	Reasoning	Safety	Avg.	Helpful	Harmless	Avg.
BT RM	93.02	57.02	84.98	77.43	78.11	74.38	73.23	73.81
FR RM	94.41	57.02	83.86	78.24	78.38	73.49	73.62	73.55
FC RM	94.41	53.29	85.53	76.76	77.50	74.30	73.62	73.96

length in Appendix C.

Data Selection We will show that this fair distribution of rewards will bring extra benefits in data selection. Figure 4 shows the performance on AlpacaEval2 when using different RMs to select samples. From the figure, we can draw two conclusions: (1) When sampling the same number of samples, FR RM and FC RM can select higher quality samples compared to BT RM. (2) To achieve the same performance, FR RM and FC RM require fewer samples, indicating higher sampling efficiency.

Social Bias We further validated our methods on the CrowS-Pairs¹ (Nangia et al., 2020) dataset, which includes sentences with social biases. This dataset contains two types of sentences: “sent more”, which is more stereotypical, and “sent less”, which is less stereotypical. It encompasses nine types of biases, such as gender and nationality. As shown in Figure 5, the BT RM tends to assign higher rewards to the more stereotypical sentences, resulting in a larger distributional difference between “sent more” and “sent less”. This indicates that the BT RM exhibits unfairness across various social biases. In contrast, FR RM and FC RM show smaller distributional differences, demonstrating greater fairness across different social biases.

These findings highlight the effectiveness of our Fair RMs in providing a fairer reward distribution

while maintaining high performance and sampling efficiency.

6.1.2 Fairness Policy Model

Table 2: Performance of different policy models on AlpacaEval2 and MT-Bench.

	AlpacaEval2		MT-Bench
	LC WR	WR	Overall
LLaMA3	SFT	14.34	8.17
	R-DPO	20.87	11.16
	KTO	19.44	16.64
	DPO	16.71	14.23
	+FR	20.48	15.74
	+FC	21.10	16.96
Qwen2.5	SFT	13.47	8.11
	R-DPO	19.95	10.15
	KTO	17.81	14.39
	DPO	18.93	13.18
	+FR	21.05	15.25
	+FC	19.72	14.53

Table 2 presents the results of different policy models on AlpacaEval2 and MT-Bench. It can be observed that our fairness reward methods, when combined with DPO, consistently demonstrates superior performance on AlpacaEval2 and MT-Bench with both LLaMA3-SFT and Qwen2.5-SFT as base models. This highlights the effectiveness of our fairness rewards method. The success of our meth-

¹<https://github.com/nyu-mll/crows-pairs/>

ods can be attributed to their ability to implicitly fit a fairness RM during the training of policy models, thereby generating higher quality outputs.

Combining Fair DPOs with Fair RM further enhances performance. We sample the policy model 1, 8, 16, 32, and 64 times, using Fair RM to select the best sample. We recorded the lengths and performance of these samples and fitted a curve to this data, as shown in Figure 6. It can be observed that for the same model, performance gradually increases with length, indicating a correlation between performance and length. However, for different models, aligned models produce higher quality outputs compared to the SFT model, but their outputs are also longer. Among the three aligned models, Fair DPOs achieve better performance than DPO while producing shorter outputs, suggesting that our model can mitigate length bias to some extent.

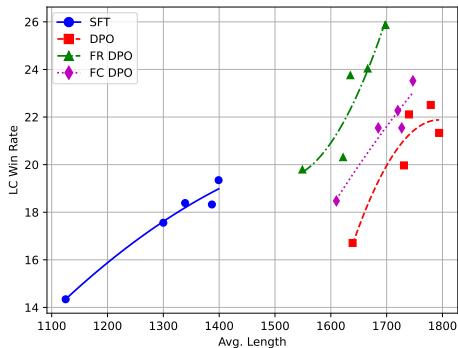


Figure 6: Length and performance relationships of samples for different models.

6.2 Ablation Study ($\mathcal{RQ}2$)

Table 3: Performance under different fairness functions.

Model	AlpacaEval2		MT-Bench Overall Score
	LC WR	WR	
DPO	16.71	14.23	6.46
FR DPO			
$\tau = -5$	19.72	14.44	6.59
$\tau = -1$	20.48	15.74	6.70
$\tau = 0.5$	20.01	15.21	6.56
$\tau = 2$	20.01	17.35	6.69
$\tau = 10$	19.98	16.24	6.62

Impact of Fairness Function Eq (7) presents a unified metric for measuring fairness, from which

different fairness functions can be derived by varying τ . We aim to explore the impact of different fairness functions on performance. We experiment with various τ values within the range of $[-5, 10]$, and the results are summarized in Table 3. It can be observed that the performance of FR DPO consistently surpasses that of the native DPO across all fairness functions. This indicates that our method is robust to variations in τ . The reason for this robustness is that all fairness functions derived from the unified metric satisfy the three desired properties, ensuring that the rewards obtained are fair.

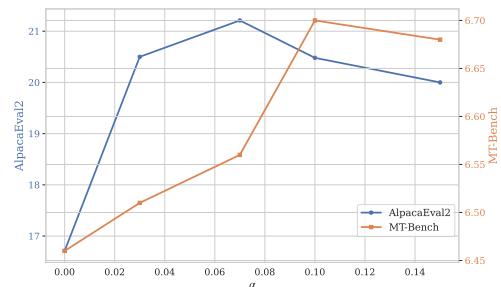


Figure 7: Performance under different fairness contribution α .

Impact of Fairness Contribution α We conduct experiments by fixing $\tau = -1$ and varying the fairness contribution α within the range of $[0, 0.15]$, as illustrated in Figure 7. When $\alpha = 0$, the model reduces to the native DPO. As α increases, the model’s performance on AlpacaEval2 and MT-Bench initially improves and then declines. This trend occurs because, at lower values of α , enhancing fairness contributes to better output quality. However, when α becomes too large, the excessive emphasis on fairness leads to a compromise in utility. Considering the trade-off between fairness and utility, we typically set $\alpha = 0.1$ in practical experiments. We present the ablation experiments on the Fairness Contribution γ in Appendix C.

7 Conclusion

In this paper, we tackle the critical issue of reward unfairness in RLHF. We identify that length bias and category bias are specific case of the broader problem of reward unfairness. To address this comprehensively, we introduce the reward fairness framework, which models preference learning as a resource allocation problem to balance fairness and utility in reward distribution. We propose two methods to achieve fairness rewards: Fairness Regularization and Fairness Coefficient. These meth-

ods are applied in two key scenarios: training a fairness RM for verification and training a policy model for reinforcement learning that implicitly incorporates fair rewards.

Limitation

We investigate the issue of rewards unfairness and proposed a solution from the perspective of resource allocation, validating our approach in both verification and RL scenarios. The limitations of this study are summarized as follows: (1) Reward unfairness is a broad concept, and this paper primarily focuses on category bias and length bias, with a simple validation on social bias. However, reward unfairness may be related to various issues in reward models, such as reward hacking. (2) Our Fairness Rewards method can seamlessly integrate with RM and RL frameworks that utilize RMs either explicitly or implicitly. We have only validated it on BT models and DPO, but the Fairness Rewards method has the potential for broader applications.

Acknowledgements

We would like to express our sincere gratitude to all the anonymous reviewers for their invaluable feedback that greatly improved this paper. In particular, special thanks to Dr. Weiran Shen for his insightful suggestions and invaluable assistance in early-stage discussions. This research was supported by National Natural Science Foundation of China (No.62476277), National Key Research and Development Program of China (NO. 2024YFE0203200), CCF-ALIMAMA TECH Kangaroo Fund(No.CCF-ALIMAMA OF 2024008), and Huawei-Renmin University joint program on Information Retrieval. We also acknowledge the support provided by the fund for building world-class universities (disciplines) of Renmin University of China and by the funds from Beijing Key Laboratory of Big Data Management and Analysis Methods, Gaoling School of Artificial Intelligence, Renmin University of China, from Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, from Intelligent Social Governance Interdisciplinary Platform, Major Innovation & Planning Interdisciplinary Platform for the “DoubleFirst Class” Initiative, Renmin University of China, from Public Policy and Decision-making Research Lab of Renmin University of China, and from Public Com-

puting Cloud, Renmin University of China.

References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Arxiv Preprint Arxiv:2204.05862*.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345.
- Lichang Chen, Chen Zhu, Juhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Odin: Disentangled reward mitigates hacking in rlhf. In *Forty-first International Conference on Machine Learning*.
- 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](#). *Transactions on Machine Learning Research*.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024a. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Arxiv Preprint Arxiv:2404.04475*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024b. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.
- 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*, 50:1097–1179.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. [Scaling laws for reward model overoptimization](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 10835–10866.
- William M Hayes, Nicolas Yax, and Stefano Palminteri. 2024. Large language models are biased reinforcement learners. *Arxiv Preprint Arxiv:2405.11422*.
- Zhengxu Hou, Bang Liu, Ruihui Zhao, Zijing Ou, Yafei Liu, Xi Chen, and Yefeng Zheng. 2021. Imperfect also deserves reward: Multi-level and sequential reward modeling for better dialog management. *Arxiv Preprint Arxiv:2104.04748*.

- Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. **Position: TrustLlm: Trustworthiness in large language models.** In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 20166–20270.
- Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. 1984. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, Ma*, 21:1.
- Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. 2017. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. In *International Conference on Machine Learning*, pages 1645–1654.
- Naoki Katoh and Toshihide Ibaraki. 1998. Resource allocation problems. *Handbook of Combinatorial Optimization: Volume 1–3*, pages 905–1006.
- Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *Arxiv Preprint Arxiv:2312.14925*.
- Sungdong Kim and Minjoon Seo. 2024. Rethinking the role of proxy rewards in language model alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20656–20674.
- A Kumar and J Kleinberg. 2000. Fairness measures for resource allocation. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 75–75.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2024. Rewardbench: Evaluating reward models for language modeling. *Arxiv Preprint Arxiv:2403.13787*.
- Tian Lan, David Kao, Mung Chiang, and Ashutosh Sabharwal. 2010. **An axiomatic theory of fairness in network resource allocation.** In *2010 Proceedings IEEE INFOCOM*, pages 1–9.
- Jiahui Li, Hanlin Zhang, Fengda Zhang, Tai-Wei Chang, Kun Kuang, Long Chen, and Jun Zhou. 2024. Optimizing language models with fair and stable reward composition in reinforcement learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10122–10140.
- Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *Arxiv Preprint Arxiv:2308.10149*.
- Wei Liu, Yang Bai, Chengcheng Han, Rongxiang Weng, Jun Xu, Xuezhi Cao, Jingang Wang, and Xunliang Cai. 2024. Length desensitization in directed preference optimization. *Arxiv Preprint Arxiv:2409.06411*.
- Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. 2024. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. *Arxiv Preprint Arxiv:2406.10957*.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Arxiv Preprint Arxiv:2405.14734*.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. 2024. Beyond the binary: Capturing diverse preferences with reward regularization. *Arxiv Preprint Arxiv:2412.03822*.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. **Disentangling length from quality in direct preference optimization.** In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussonot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024a. Warp: On the benefits of weight averaged rewarded policies. *Arxiv Preprint Arxiv:2406.16768*.

- Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024b. Warm: On the benefits of weight averaged reward models. *Arxiv Preprint Arxiv:2401.12187*.
- David Reber, Sean Richardson, Todd Nief, Cristina Garbacea, and Victor Veitch. 2024. Rate: Score reward models with imperfect rewrites of rewrites. *Arxiv Preprint Arxiv:2410.11348*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Arxiv Preprint Arxiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Arxiv Preprint Arxiv:2402.03300*.
- Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. **Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback.** In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2859–2873.
- Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.
- Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *Arxiv Preprint Arxiv:2409.11239*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.
- Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *Arxiv Preprint Arxiv:2406.12845*.
- Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024b. Helpsteer2-preference: Complementing ratings with preferences. *Arxiv Preprint Arxiv:2410.01257*.
- Yuzi Yan, Xingzhou Lou, Jialian Li, Yiping Zhang, Jian Xie, Chao Yu, Yu Wang, Dong Yan, and Yuan Shen. 2024. Reward-robust rlhf in llms. *Arxiv Preprint Arxiv:2409.15360*.
- Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. 2024. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment.
- In *Forty-first International Conference on Machine Learning*.
- Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *Arxiv Preprint Arxiv:2408.15240*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Further Discussions

Further explanation of Figure 1 Someone may argue that “ y_1^l could be a better response than y_2^w ”. Our objective is to achieve both utility and fairness in reward allocation in RLHF. Figure 2 in our paper effectively illustrates our objective. Considering an example where four responses (y_1, y_2, y_3 , and y_4) are sampled for the same prompt, with rewards ranked as $y_1 > y_2 > y_3 > y_4$. For preference data pairs (y_1, y_2) and (y_3, y_4), it is reasonable that $y_2 > y_3$. This does not conflict with our objective, as utility measures the alignment of reward allocation with human preferences, meaning that a good response should receive a higher reward than a poor one. Fairness focuses on the reward distribution across different types of samples, typically from various domains. These absolute reward values are generally incomparable, but we expect their distributions to be as fair as possible to avoid issues in downstream scenarios such as rejected sampling and RL. Unfortunately, our experiments reveal that the commonly used BT model exhibits pervasive reward unfairness, as shown in Figure 3 and Figure 8. This unfairness affects both in-distribution and out-of-distribution data, leading to category and length biases. Additional, Table 4 presents the average rewards from the Bradley-Terry (BT) model on the HH-RLHF dataset. For both “helpful” and “harmless” data, the rewards for “chosen” are greater than those for “rejected”, aligning with the utility objective. However, the rewards for “helpful” are significantly higher than those for “harmless”, which is unfair. When such unfair rewards are used in rejected sampling and RL, the model’s output becomes more helpful but neglects harmlessness.

Table 4: Average rewards of BT model on the HH-RLHF dataset.

	Helpful		Harmless	
	chosen	rejected	chosen	rejected
Avg. Reward	-1.39	-2.26	-4.15	-5.23

Fairness and Utility Figure 2 illustrates our dual objectives: achieving fairness and utility in reward allocation. The relationship between these objectives varies slightly between verification and reinforcement learning scenarios. In the verification scenario, “fairness” aims to make the distribution of different types of rewards more consistent, while “utility” ensures that for any given prompt, the reward for a good response exceeds that for a bad response. These objectives are inherently independent and non-conflicting, though balancing them necessitates a multi-objective optimization. In the reinforcement learning scenario, “fairness” can even enhance “utility” by guiding the model’s output to be both more helpful and harmless.

B Experiment Setting

Datasets & Baselines In the verification scenario, we conduct experiments on two benchmarks: Reward Bench (Lambert et al., 2024) and HH-RLHF (Bai et al., 2022). Our RMs are trained using the training set from HH-RLHF², making HH-RLHF an in-distribution (ID) benchmark, while Reward Bench serves as an out-of-distribution (OOD) benchmark. We report the accuracy on both benchmarks. For the RL scenario, we evaluate our methods on the AlpacaEval2 (Dubois et al., 2024a) and MT-Bench (Zheng et al., 2023) benchmarks. We provide results for AlpacaEval2 in terms of Length-controlled Win Rate (LC WR) and Win Rate (WR), and for MT-Bench, we report the overall score. For the policy model, we utilize the UltraFeedback Binarized³ and SHP⁴ datasets for training. We train the policy model using different methods such as DPO (Rafailov et al., 2024), KTO (Ethayarajh et al., 2024) and R-DPO (Park et al., 2024) with HALOs⁵.

²<https://huggingface.co/datasets/Anthropic/hh-rlhf>

³https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

⁴<https://huggingface.co/datasets/stanfordnlp/SHP>

⁵<https://github.com/ContextualAI/HALOs/>

Training Setting For the reward model, we train on the HH-RLHF training set for one epoch with a learning rate of 2×10^{-6} and a global batch size of 256. For the policy model, we utilize the UltraFeedback Binarized and SHP datasets, training for one epoch with a learning rate of 5×10^{-6} and a global batch size of 256. During sampling with the policy model, the temperature coefficient is set to 1. All experiments are performed on an $8 \times$ H800 machine. Both the reward models and the policy models are trained using LLaMA3-SFT⁶ (a base model developed by Dong et al. (2024)) and Qwen2.5-SFT (a base model we trained⁷ following Dong et al. (2024)). Qwen2.5-SFT is a model we trained based on the Qwen2.5-7B base using the dataset from RLHFow⁸ for one epoch. The global batch size was set to 128, and the learning rate was 2×10^{-5} . For all policy models, the β parameter was uniformly set to 0.1. Additionally, the desirable weight and undesirable weight for the KTO were both set to 1.

C Supplement Experiment

Rewards on Length The rewards distribution of BT RM, FR RM, and FC RM across different lengths on the HH-RLHF dataset is illustrated in Figure 8. Our FR RM and FC RM exhibits a more consistent rewards distribution across varying lengths, demonstrating that our method effectively mitigates length bias.

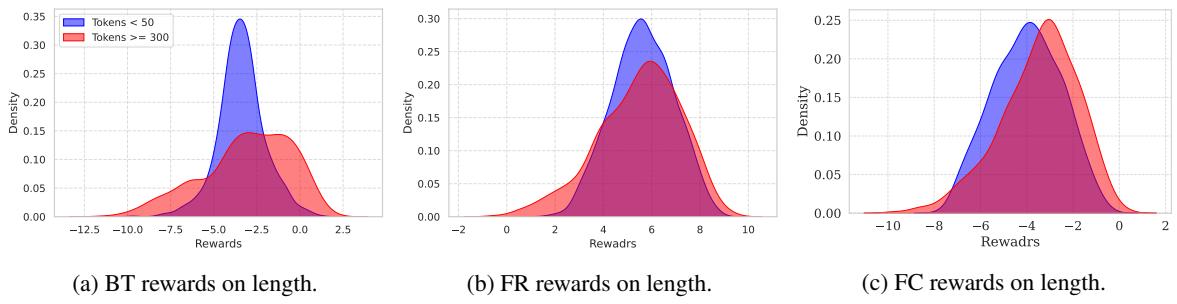


Figure 8: Rewards on Length

Impact of Fairness Contribution γ We conduct experiments by fixing $\tau = -1$ and varying the fairness contribution γ within the range of $[0, 1.5]$, as illustrated in Figure 9. We obtain conclusions similar to those with fairness contribution α . Considering the trade-off between fairness and utility, we typically set $\gamma = 0.5$ in practical experiments.

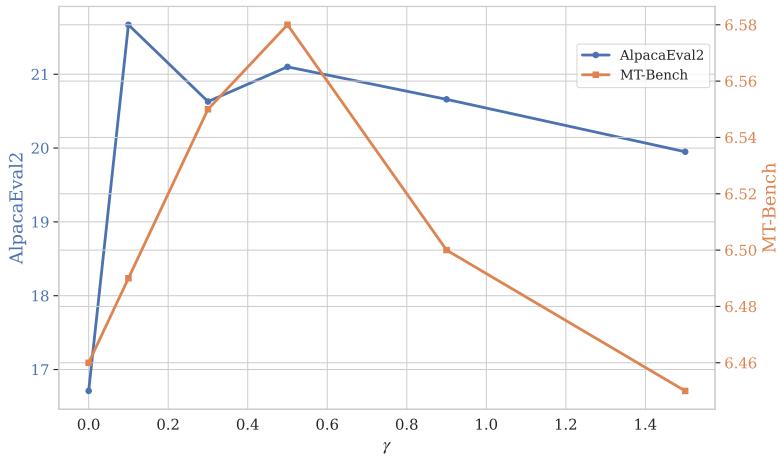


Figure 9: Performance under different fairness contribution γ .

⁶<https://huggingface.co/RLHFow/LLaMA3-SFT>

⁷<https://github.com/RLHFow/Online-RLHF/>

⁸<https://huggingface.co/datasets/RLHFow/RLHFow-SFT-Dataset-ver2>