CHARM: Calibrating Reward Models With Chatbot Arena Scores

Xiao Zhu[*,1], **Chenmien Tan**[*,2], **Pinzhen Chen**[3], **Rico Sennrich**[4], **Yanlin Zhang**[1], **Hanxu Hu**[†,4]

[1]HKUST (Guangzhou), [2]Alibaba Group, [3]University of Edinburgh, [4]University of Zurich

Abstract

Reward models (RMs) play a crucial role in Reinforcement Learning from Human Feedback by serving as proxies for human preferences in aligning large language models. In this paper, we identify a model preference bias in RMs, where they systematically assign disproportionately high scores to responses from certain policy models. This bias distorts ranking evaluations and leads to unfair judgments. To address this issue, we propose a calibration method named **CH**atbot **A**rena calibrated **R**eward **M**odeling (**CHARM**) that leverages Elo scores from the Chatbot Arena leaderboard to mitigate RM overvaluation. We also introduce a Mismatch Degree metric to measure this preference bias. Our approach is computationally efficient, requiring only a small preference dataset for continued training of the RM. We conduct extensive experiments on reward model benchmarks and human preference alignment. Results demonstrate that our calibrated RMs (1) achieve improved evaluation accuracy on RM-Bench and the Chat-Hard domain of RewardBench, and (2) exhibit a stronger correlation with human preferences by producing scores more closely aligned with Elo rankings. By mitigating model preference bias, our method provides a generalizable and efficient solution for building fairer and more reliable reward models.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF; Ouyang et al., 2022; Christiano et al., 2017) has emerged as a fundamental approach for aligning large language models (LLMs) with human values, ensuring they generate helpful, coherent, and safe responses (Achiam et al., 2023; Touvron et al., 2023; Gemini Team et al., 2023; Bai et al., 2023). At the core of RLHF are reward models (RMs). RMs are typically trained on pairwise preference data, where human annotators evaluate multiple model-generated responses and rank them based on specific criteria (Ouyang et al., 2022; Lee et al., 2024). Given these ranked preferences, the RM learns to predict which responses humans would favor, effectively acting as an automated judge in place of human raters. Beyond RMs, LLM-as-a-Judge systems (Li et al., 2023) have emerged as an alternative to human evaluation, where pretrained LLMs are used as evaluation models to score and rank model-generated responses, replacing human feedback at a lower cost.

Nonetheless, both RMs and LLM-as-a-Judge can suffer from biases that compromise the fairness of evaluations. Recently, several studies have focused on detecting bias in RMs and LLM-as-a-judge systems. Park et al. (2024) and Ye et al. (2024) identified various types of bias, highlighting the pervasive nature of bias in evaluation models. These biases often manifest as a preference for specific answer-related patterns, such as length bias or position bias. However, some biases like self-preference bias, where judge models tend to favor outputs

---

*Equal contribution. †Correspondence to hanxu.hu@uzh.ch.
Our code is available at https://github.com/HexagonStar/CHARM.

generated by themselves or similar LLMs, are more subtle and difficult to identify. When biased, RMs can be easily exploited, where policy models optimize for reward in ways that deviate from genuine human preferences (Eisenstein et al., 2023; Gao et al., 2023; Pang et al., 2022). This phenomenon, referred to as reward hacking (Skalse et al., 2022), allows models to exploit vulnerabilities in the RM rather than achieving genuine behavioral improvements, potentially leading to deceptive or unintended outcomes. Similarly, bias in LLM-as-a-Judge systems weakens the correlation between LLM judges and human annotators, leading to inaccurate evaluations and unfair model rankings (Dubois et al., 2024).

In this paper, we identify a model preference bias in popular RMs, where they systematically give disproportionately high scores to certain policy models beyond what is justified by human preferences. This bias can distort model rankings, leading to unfair advantages in RLHF training and evaluation. Fortunately, over-valued policy models can naturally serve as generators of false positives, providing valuable counterexamples for debiasing. To address this issue, we propose a calibration method named **CH**atbot **A**rena calibrated **R**eward **M**odeling (**CHARM**). By selecting the over-valued and reference model pair, we construct debiased preference pairs leveraging Elo scores from Chatbot Arena (Chiang et al., 2024) to correct this preference bias and improve the RM's alignment with human judgment.

To summarize, the major contributions of our work are as follows:

- We uncover model preference bias in RMs, where certain policy models are over-valued systematically. To measure this bias, we introduce a Mismatch Degree metric, quantifying the misalignment between over-valued models and human preferences.

- We propose a calibration method **CHARM** that leverages counterpart model pairs and their Elo scores from the Chatbot Arena leaderboard. Compared with previous calibration methods, our approach is computationally efficient, requiring only a small preference dataset and arena scores information from a limited set of models.

- We conduct extensive experiments across reward model benchmarks and additional tasks, demonstrating that our calibration method effectively mitigates bias and improves alignment with human preferences.

2 Related Work

LLM-as-a-judge Both LLM-as-a-judge and reward models are widely used in LLM evaluation and preference learning. For evaluation, a series of benchmarks such as MT-Bench (Zheng et al., 2023), Alpaca-Eval (Dubois et al., 2023), and Arena-hard (Li et al., 2024) are used to evaluate the quality of the model's responses. A series of works, such as Ultrafeedback (Cui et al., 2024) and RLAIF (Lee et al., 2024), use LLMs for preference annotation which correlate training signals with evaluation measures.

Bias Mitigation LLM-as-a-judge might bring biases and affect the accuracy of the evaluation. Recently, several studies have focused on detecting and mitigating bias in RMs and LLM-as-a-judge. Park et al. (2024) identified six distinct types of bias in evaluation models and leveraged LLMs to construct a debiased dataset. A notable example of bias arises when models exploit judges' preferences for specific characteristics (e.g., response length), leading to inflated performance estimations. To address this, Dubois et al. (2024) proposed a regression-based method to mitigate length bias, while Huang et al. (2025) introduced a post hoc calibration technique for reward models. Beyond these biases, Li et al. (2025) found that judge models may develop model preference bias, favoring content generated by themselves or closely related LLMs due to their exposure to synthetic data.

Reward Models Evaluation Recently, Lambert et al. (2024) curated RewardBench to evaluate the performance of reward models by letting RM select the better response from a given pair and calculate the accuracy. RM-Bench (Liu et al., 2025) evaluates reward models based on their sensitivity to subtle content differences and resistance to style biases.
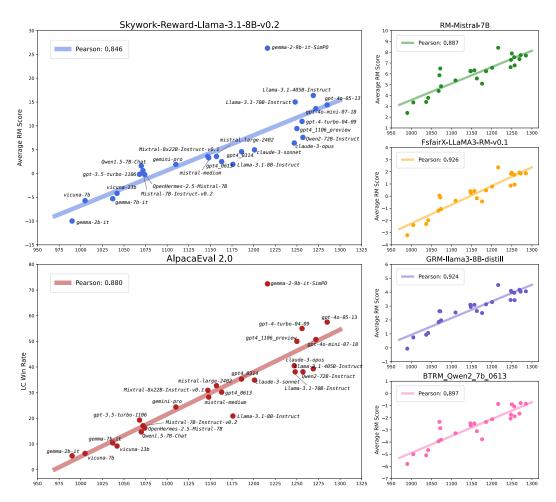
Figure 1: Average reward model scores across policy models on AlpacaEval. The x-axis represents arena Elo scores. The left lower plot illustrates the Length-Controlled win rates of these models on AlpacaEval.

3 Preference Bias in Reward Models

Reward models are designed with the primary goal of aligning AI-generated responses with human preferences. One of the most direct reflections of human preferences in large-scale AI evaluation is Chatbot Arena (Zheng et al., 2023), where real users interact with language models and rank them based on the content they generate. The Elo scores derived from Chatbot Arena provide a robust measure of how well models align with human preferences.

Given this, we make the assumption that an ideal RM should produce scores that closely correlate with Chatbot Arena's Elo scores. To validate this assumption, we use AlpacaEval (Li et al., 2023) as our evaluation dataset, which consists of 805 carefully curated questions. This dataset has been shown to exhibit a 98% Spearman correlation with the evaluations from Chatbot Arena.

We select five widely used reward models (Liu et al., 2024; Dong et al., 2023; Xiong et al., 2024; Yang et al., 2024), and a diverse set of policy models with varying Arena Elo scores. For each policy model, we use each reward model to score all responses and take the average. We also display the Length-Controlled win rates of these policy models on AlpacaEval (Dubois et al., 2024). The results are shown in Figure 1:

Observation 1. RM Scores Correlate Positively with Human Preferences From Figure 1, we observe that models with higher Elo scores in Chatbot Arena tend to receive higher RM scores on their responses. This supports our initial assumption that an ideal RM should

reflect human preference rankings. We compute the Pearson correlation between policy models' RM and Elo scores for each RM (Also see Figure 1). The results indicate a strong positive correlation across all tested RMs.

Observation 2. While RM Scores Align Globally, Local Misalignments Exist Although RM scores exhibit an overall alignment with human preferences, they sometimes deviate for specific policy models, assigning scores that are inconsistent with the models' Elo scores.

Observation 3. RMs May Favor Certain Policy Models Unfairly Some policy models, such as Gemma-2-9b-it-SimPO (Meng et al., 2024), receive disproportionately high RM scores, sometimes even surpassing significantly stronger models based on Elo rankings (More results are in Appendix A.1). A similar trend is observed in the AlpacaEval leaderboard, suggesting that this bias may also exist in LLM-as-a-judge systems. Notably, the over-valued models share a similarity in that they undergo preference optimization, particularly when LLM-generated annotations are involved, which may introduce systematic biases.

We observe that reward models exhibit a form of model preference bias, overestimating certain policy models by assigning them scores higher than expected. This bias may stem from optimization techniques that favor models trained on LLM-annotated data. While its exact origins may vary, the impact remains consistent: reward models fail to provide fair evaluations. Therefore, the paper proposes a method to correct mis-calibrated reward models that may over-value policy models or responses, as we introduce in the next section.

4 Methodology

4.1 Preliminaries

Reward Modeling A reward model assigns scores to responses generated by large language models, helping rank and select the most human-aligned outputs. Formally, let $\mathcal{D} = \{(x, y)\}$ represent a dataset of instruction-response pairs, where $x \in \mathcal{X}$ is an instruction and $y \in \mathcal{Y}$ is a response. A reward model $r_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ predicts the score $r_\phi(x, y)$ for a response y conditioned on an instruction x.

Most RMs are trained using pairwise preference data, which consists of triplets (x, y^+, y^-), where y^+ is the preferred response over y^-. The RM is trained to optimize a Bradley-Terry pairwise ranking loss (Ouyang et al., 2022) with $\sigma(\cdot)$ being a sigmoid function:

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\log\sigma\left(r_\phi(x, y^+) - r_\phi(x, y^-)\right)\right] \tag{1}$$

4.2 CHARM: Chatbot Arena Calibrated Reward Modeling

Given a set of instructions $\mathcal{X} = \{x_i\}_{i=1}^N$ and two policy models, one over-valued model π_O and one reference model π_R, for each instruction x_i, the two models generate responses $y_i^O \sim \pi_O$ and $y_i^R \sim \pi_R$. The reward model r_ϕ assigns scores to these responses, producing $s_i^O = r_\phi(x_i, y_i^O)$ and $s_i^R = r_\phi(x_i, y_i^R)$, resulting in a preference dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-) \mid y_i^+ = \arg\max(s_i^O, s_i^R)\}_{i=1}^N$. Since the reward model r_ϕ overestimates the responses from model π_O, the resulting preference dataset $\mathcal{D}$ inherits this bias. To address this issue, **CHARM** reconstructs a debiased preference dataset to mitigate the preference bias in reward modeling.

The Elo rating system (Elo, 1967) provides a probabilistic model for ranking players (or models, in this case) based on their relative performance. Following Chatbot Arena's implementation where a model gets a score of 1 for a win, 0.5 for a tie, and 0 for a loss, for an over-valued model with Elo_O and a reference model with Elo_R, the expected win rate of the over-valued model $\mathbb{P}(O)$ is defined as Equation 2. $\mathbb{P}(O)$ can also be expressed as a weighted sum of the probabilities of the over-valued model getting a win $\mathbb{P}_{win}$ and a tie $\mathbb{P}_{tie}$:

$$\mathbb{P}(O) = \frac{1}{1 + 10^{(\text{Elo}_R - \text{Elo}_O)/400}} = \mathbb{P}_{win} + 0.5\mathbb{P}_{tie} \tag{2}$$

Ties in RM are rare unless both models produce identical responses, naturally requiring $\mathbb{P}_{tie} \to 0$. Nonetheless, Chatbot Arena's scoring implementation (1/0.5/0 for win/tie/loss) allows us to evenly split $\mathbb{P}_{tie}$ between wins and losses, maintaining equivalent Elo scores. Therefore, our Elo-derived win rate can be directly applicable to a strict win/loss scenario for RMs. Given the observation in section 3 that RM scores are correlated with Elo scores, we contend that if the RM were perfectly aligned with human preferences, its empirical win rate should match this probability:

$$\hat{\mathbb{P}}(O) = \frac{1}{N} \sum_{i=1}^{N} \sigma(s_i^O - s_i^R) \approx \mathbb{P}(O) \tag{3}$$

However, in practice, we find that there exist deviations between $\hat{\mathbb{P}}(O)$ and $\mathbb{P}(O)$ because of model preference bias. To correct this bias, we seek a transformation of RM scores such that the empirical win rate $\hat{\mathbb{P}}'(O)$ after calibration better aligns with the expected win probability $\mathbb{P}(O)$. We introduce a score offset Δ applied to the RM scores of over-valued policy model's responses: $s_i'^O = s_i^O + \Delta$, then the calibrated empirical win rate will be: $\hat{\mathbb{P}}'(O) = \frac{1}{N} \sum_{i=1}^{N} \sigma(s_i'^O - s_i^R)$. Our goal is to find Δ that minimizes the deviation from the theoretical probability. We optimize Δ by minimizing the MSE loss:

$$\mathcal{L}(\Delta) = \mathbf{MSE}\left(\frac{1}{N} \sum_{i=1}^{N} \sigma(s_i^O + \Delta - s_i^R), \mathbb{P}(O)\right) \tag{4}$$

After determining the offset Δ, we can construct a calibrated preference dataset for further reward model training. See Algorithm 1 for the detailed pseudo-code implementation.

Algorithm 1: CHARM: CHatbot Arena calibrated Reward Modeling

Input: Instruction set $\mathcal{X} = \{x_i\}_{i=1}^{N}$, over-valued model π_O, reference model π_R, Elo scores Elo_O and Elo_R, reward model r_ϕ, epochs T
Output: Calibrated preference dataset $\mathcal{D}'$
for *each instruction $x_i \in \mathcal{X}$* **do**
    Generate responses: $y_i^O \sim \pi_O$, $y_i^R \sim \pi_R$;
    Compute RM scores: $s_i^O = r_\phi(x_i, y_i^O)$, $s_i^R = r_\phi(x_i, y_i^R)$;
Compute expected win probability $\mathbb{P}(O)$ as Equation 2;
Initialize Δ;
for $t = 1$ *to* T **do**
    Compute $\hat{\mathbb{P}}'(O)$ as Equation 3;
    Compute loss $\mathcal{L}$ as Equation 4;
    Update Δ;
for *each instruction $x_i \in \mathcal{X}$* **do**
    Apply score offset: $s_i'^O = s_i^O + \Delta$;
Construct calibrated preference dataset:

$$\mathcal{D}' = \{(x_i, y_i^+, y_i^-) \mid y_i^+ = \arg\max(s_i'^O, s_i^R)\}_{i=1}^{N}$$

    return $\mathcal{D}'$;

4.3 A Metric for Model Preference Bias Measurement

To quantify the misalignment between a reward model's preference for different policy models and human preferences, we introduce a Mismatch Degree metric. This metric shows the discrepancy between the reward model's scoring and the expected human preference reflected by Elo scores, measuring the degree of RM's model preference bias.

Given a model π_O, a reference model π_R, and a preference dataset built upon them, we define the Mismatch Degree (MD) between them as:

$$\mathbf{MD}(\pi_O, \pi_R) = \left| \frac{\hat{\mathbb{P}}(O) - \mathbb{P}(O)}{\max(\mathbb{P}(O), 1 - \mathbb{P}(O))} \right| \tag{5}$$

where $\hat{\mathbb{P}}(O)$ is the probability of model π_O winning against π_R according to the reward model's scores. $\mathbb{P}(O)$ is the expected win rate of π_O over π_R, derived from their Elo scores in Chatbot Arena. This metric captures how much the reward model's judgments deviate from the expected human preference. A positive $\hat{\mathbb{P}}(O) - \mathbb{P}(O)$ indicates that the reward model over-values model π_O relative to what is expected from human preferences while a negative value indicates an under-value.

5 Experiments

5.1 Experiment Setup

5.1.1 Experiment Design

In this section, we aim to address the following questions through experiments to validate the effectiveness of our calibration method:

Question 1. Does calibration enhance the reward model's judging capability, leading to more accurate and reliable evaluations? How does the mismatch degree influence calibration effectiveness, and is there a measurable correlation?

• We evaluate calibrated reward models on benchmarks such as RM-Bench and Reward-Bench, which consist of diverse instructions paired with two candidate responses. The reward model must assess and select the better response, providing a robust framework to measure its judging capability. We analyze the impact of MD on calibration performance by selecting model pairs with varying MD values as over-valued and reference models. By examining how calibration affects performance across these pairs, we explore the correlation between initial misalignment and calibration effectiveness.

Question 2. Does calibration successfully reduce preference bias, improving alignment with human preferences?

• We construct a battlefield using responses from various LLMs on 805 prompts from AlpacaEval. Each response is scored by the reward models, allowing us to compute pairwise win rates between models. We then compare these RM-derived win rates against Elo-based win rates obtained from Chatbot Arena, which reflect human preferences. By analyzing their alignment, we assess whether the calibration process effectively reduces bias and produces rankings that more accurately reflect human judgments.

5.1.2 Implementation Details

For preference dataset construction, we use Preference700K (Dong et al., 2024), a comprehensive dataset that aggregates preference data from eight sources. We randomly sampled 20K instructions from Preference700K and generated corresponding responses using selected over-valued and reference models. These responses were then scored by a reward model. This process produced the uncalibrated preference dataset, which served as the foundation for applying our proposed method to construct the calibrated preference dataset.

We set temperature $\tau = 0.7$ and Top_p $= 0.9$ during inference. We selected five reward models for calibration: Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024), RM-Mistral-7B, FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023; Xiong et al., 2024), GRM-llama3-8B-distill (Yang et al., 2024), and BTRM-Qwen2-7b-0613. During reward model fine-tuning, we used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 2e-6, a weight decay of 0.001, and a cosine learning rate scheduler. The models were trained for 1 epoch.

| Reward Models | Mismatch Degree | RM-Bench | | | | | | | | RewardBench |
		Chat	Math	Code	Safety	Hard	Normal	Easy	*Avg*	Chat-Hard
Skywork-RM		68.7	62.0	52.8	95.9	47.5	73.7	88.4	69.9	88.8
w/o calibration	0.639	68.9	61.9	53.1	95.9	47.6	73.8	88.5	70.0	88.8
w/ calibration		73.9	62.4	53.9	95.8	49.3	75.8	89.4	71.5	89.4
FsfairX-RM		62.5	63.2	54.6	90.4	44.9	71.6	86.5	67.7	65.3
w/o calibration	0.554	63.1	63.4	53.6	90.4	45.8	71.6	85.4	67.6	65.1
w/ calibration		64.5	63.3	56.0	90.0	45.4	72.6	87.5	68.5	65.7
Mistral-RM		60.8	56.6	52.6	88.7	37.5	68.2	88.3	64.7	60.5
w/o calibration	0.528	61.4	57.4	53.0	88.9	40.0	68.8	86.7	65.2	62.5
w/ calibration		63.2	57.0	52.4	88.3	36.3	69.8	89.5	65.2	65.1
GRM-RM		63.6	62.0	56.9	89.1	49.6	71.8	82.2	67.9	68.4
w/o calibration	0.508	63.6	62.4	58.3	89.5	49.8	72.7	82.9	68.4	68.8
w/ calibration		66.2	62.6	58.0	89.3	48.3	73.9	84.9	69.0	68.9
BTRM-RM		60.0	61.3	53.8	89.9	37.1	71.0	90.7	66.3	58.1
w/o calibration	0.162	58.5	61.5	54.1	89.1	35.3	70.8	91.3	65.8	58.7
w/ calibration		60.2	60.5	53.8	89.6	34.8	71.1	92.1	66.0	57.8

Table 1: Results of the three versions of each reward model on the benchmarks.

5.2 Experiment Results

5.2.1 Results on Reward Model Benchmarks

Given the observations in Section 3, we select Gemma-2-9b-it-SimPO as the over-valued model and GPT-4o-mini-2024-07-18 as the reference model.

We choose five reward models from the RM-Bench leaderboard, each exhibiting varying levels of performance. We compute their MD on the selected model pair, revealing distinct deviations in how they value Gemma-2-9b-it-SimPO relative to human preferences. These reward models serve as the base models for our experiments. Following the methodology described in Section 5.1.2, we construct both uncalibrated and calibrated preference datasets for each reward model.

We evaluated three versions of each reward model on the benchmark: **(1)** the original reward model, **(2)** the reward model trained on the uncalibrated dataset, and **(3)** the reward model trained on the calibrated dataset. We select RM-Bench and RewardBench as our test benchmarks. For RewardBench, only the more challenging Chat-Hard domain is reported since other domains have shown nearly saturated results. The overall results are displayed in Table 1.

From the benchmark results across different versions of the reward model, we can summarize the following findings:

Finding 1. Biased Preference Datasets Lead to Minimal Performance Gains   Across all evaluated models, uncalibrated training led to minimal or no improvement over the original reward model. While an uncalibrated preference dataset introduces additional preference data, it does not explicitly correct biases. The underlying issues in the RM's decision boundaries remain unaddressed, resulting in no meaningful shift in performance.

Finding 2. CHARM Enhances Overall Performance   Training on the calibrated preference dataset enhances reward model performance across benchmarks. On average, RM-Bench scores improved by +0.74 points, with Skywork-RM showing the largest gain of +1.6 points.

Finding 3. CHARM Mitigates Bias and Improves Chat Evaluation Performance   Among all evaluated tasks, Chat performance saw the most substantial improvement after calibration. Skywork-RM achieved the largest gain of +5.2 points, followed by Mistral-RM of +2.4 points and FsfairX-RM of +2.0 points. And a similar trend was observed in RewardBench Chat-Hard. We attribute this improvement to calibration reducing the reward model's bias toward specific response patterns. Uncalibrated models may over-prefer certain stylistic or

Policy Models	Mismatch Degree	RM-Bench					RewardBench
		Chat	Math	Code	Safety	*Avg*	Chat-Hard
Original Skywork-RM		68.7	62.0	52.8	95.9	69.9	88.8
gemma-2-9b-it-SimPO	0.639	73.9	62.4	53.9	95.8	71.5	89.4
gemma-2-27b-it	0.225	70.7	61.9	52.9	96.7	70.5	89.2
gemma-2-9b-it	0.155	70.9	62.2	52.9	96.6	70.7	89.2
Qwen2.5-72B-Instruct	0.088	70.5	62.1	52.6	96.2	70.4	89.0
Llama-3.1-70B-Instruct	0.048	68.9	61.9	53.0	96.0	70.0	88.5
Llama-3.1-8B-Instruct	0.032	68.8	61.6	52.5	96.2	69.8	88.8

Table 2: Impact of Mismatch Degree on calibration effectiveness.

structural features of certain models, leading to skewed evaluations. By correcting these biases, calibration ensures that models assess responses more fairly and in alignment with human preferences, ultimately enhancing their reliability across different dialogue scenarios.

Additionally, we observe a potential correlation between Mismatch Degree and performance improvement after calibration. Notably, Skywork-RM, which exhibited the highest MD, achieved the most significant performance gains. In contrast, BTRM, which had the lowest MD, even experienced a slight performance degradation.

To further investigate the relationship between MD and calibration effectiveness, we design additional experiments to analyze how MD influences the impact of reward model calibration. We fix Skywork-RM as the reward model and GPT-4o-mini-2024-07-18 as the reference model. To further analyze the impact of MD on calibration effectiveness, we select multiple policy models with varying MD values and repeat the previous experiments on these model pairs. The results are presented in Table 2.

Finding 4. Mismatch Degree Serves as an Indicator of Calibration Need By analyzing the results, we observe that MD serves as a strong indicator of a model's misalignment and the potential benefits of calibration. Models with higher MD values tend to exhibit greater improvements after calibration. For instance, Gemma-2-9b-it-SimPO (MD = 0.639) benefits the most from calibration, showing significant performance gains. Conversely, models with near-zero MD, such as Qwen2.5-72B-Instruct (MD = 0.088) and Llama-3.1-8B-Instruct (MD = 0.032), experience minimal or even negative performance changes after calibration. This finding highlights that if a model is already well-aligned with human preferences, additional calibration may have little effect or even introduce instability. More importantly, it validates our proposed MD metric as a practical tool for diagnosing mis-calibration in reward models.

5.2.2 Results on Human Preference Alignment

One of the primary objectives of our calibration method is to better align the reward model's judgment with human preferences by mitigating model preference bias. To validate its effectiveness, we select 24 policy models and their responses from the AlpacaEval dataset. We then use both uncalibrated and calibrated Skywork-RM to score these responses and conduct pairwise comparisons. The results are presented in Figure 2.

Finding 5. CHARM Reduces Model Preference Bias From the win rate comparison, we observe that models' performance against GPT-4o-mini-2024-07-18 and Gemma-2-9b-it-SimPO exhibits stronger alignment with ideal human preferences. Specifically, the win rates derived from RM scores are now closer to those based on Elo scores. This indicates that calibration effectively mitigates model preference bias, enabling the reward model to provide fairer and more accurate evaluations across different responses. Additionally, we compute the MD across different models. The results reveal a clear reduction in MD, further demonstrating the improved alignment of the calibrated reward models with human preferences and a decrease in model preference bias.

Finding 6. CHARM Generalizes to Unseen Models Additionally, we selected Qwen2-72B-Instruct and mistral-large-2402, two policy models that were not used during the calibration
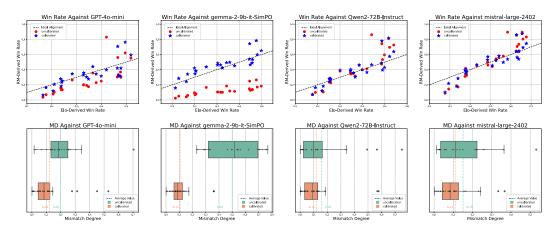
Figure 2: Win rates and Mismatch Degrees before and after calibration. In the win rate plots, the x-axis is the expected win rates calculated based on the models' Elo scores, while the y-axis is the win rates derived from the reward model scores. Points closer to the dotted line indicate a better alignment between the reward model and human preferences.

process, to evaluate whether our method generalizes to unseen LLMs. Results in Figure 2 (right) indicate that the calibrated reward model maintains a stronger correlation with human preferences even on these unseen models, further validating the generalization and effectiveness of our method.

6 Discussions

While our method focuses on discriminative RMs based on the Bradley-Terry model, other training objectives, such as focal loss (Lin et al., 2017) and hinge loss (Shawe-Taylor & Cristianini, 2004), have been explored in prior work (Liu et al., 2024). Additionally, alternative RM formulations exist, including pairwise (Jiang et al., 2023) and generative (Zhang et al., 2024) reward modeling. Our method focuses on constructing a debiased preference dataset rather than relying on the specific architecture of the reward model.

A crucial direction for future research is evaluating how calibrated reward models impact the training of policy models in RLHF and RLAIF. Since RLAIF directly leverages outputs from LLM judges instead of human-labeled preference datasets, biased reward models can propagate errors into the optimization process. Calibrating RMs could mitigate these biases, leading to more reliable reward signals and, consequently, better-aligned policy models.

In the future, we aim to further investigate the underlying mechanisms of model preference bias and plan to extend our calibration method to other types of reward models and explore its applications in broader RLHF and RLAIF scenarios.

7 Conclusion

In this paper, we study model preference bias in reward models, where RMs tend to favor responses generated by certain policy models, particularly those that have undergone preference optimization. To quantify this bias, we introduce a Mismatch Degree metric and propose a calibration method **CHARM**, which leverages Elo scores from the Chatbot Arena leaderboard to mitigate this bias. Extensive experiments demonstrate that **CHARM** enhances RMs' judging capabilities, particularly in the chat domain, leading to better alignment with human preferences and improved generalization to unseen LLMs. Further experiments confirm that Mismatch Degree serves as a reliable bias indicator, exhibiting a strong positive correlation with calibration performance. Our findings highlight previously overlooked biases in RMs, underscoring the need for further research into their underlying mechanisms.

References

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2024.

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

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf, 2024.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo Ponti, and Ivan Titov. Post-hoc reward calibration: A case study on length bias. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. RLAIF: Scaling reinforcement learning from human feedback with AI feedback, 2024.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2501.01534*, 2025.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arenahard and benchbuilder pipeline, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-bench: Benchmarking reward models of language models with subtlety and style. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Richard Yuanzhe Pang, Vishakh Padmakumar, Thibault Sellam, Ankur P Parikh, and He He. Reward gaming in conditional text generation. *arXiv preprint arXiv:2211.08714*, 2022.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. Offsetbias: Leveraging debiased data for tuning evaluators, 2024.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35: 9460–9471, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*, 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Aviral Kumar Mehran Kazemi, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

A Appendix

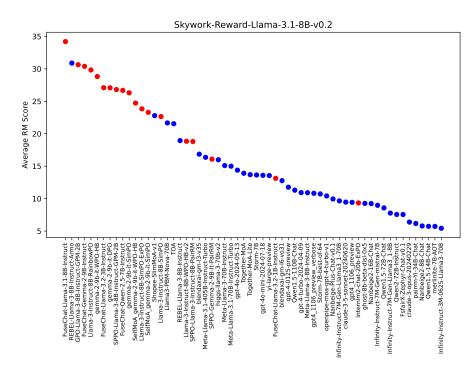## A.1 Extra Experiments on Preference Bias

Figure 3: Score results of Skywork-RM on more policy models in the AlpacaEval dataset.
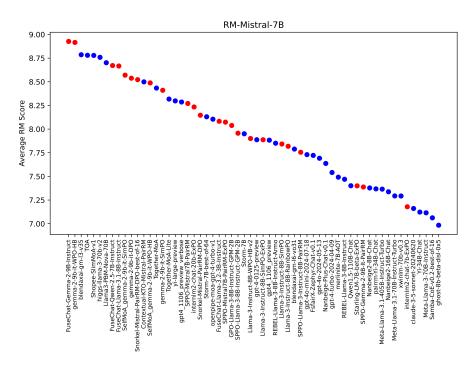

Figure 4: Score results of Mistral-RM on more policy models in the AlpacaEval dataset.
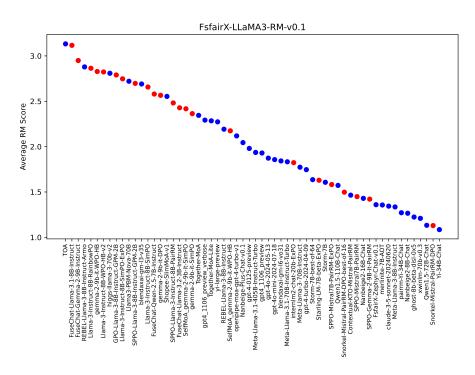
Figure 5: Score results of FsfairX-RM on more policy models in the AlpacaEval dataset.
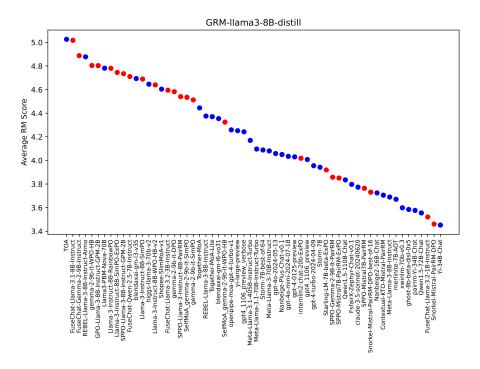

Figure 6: Score results of GRM-RM on more policy models in the AlpacaEval dataset.
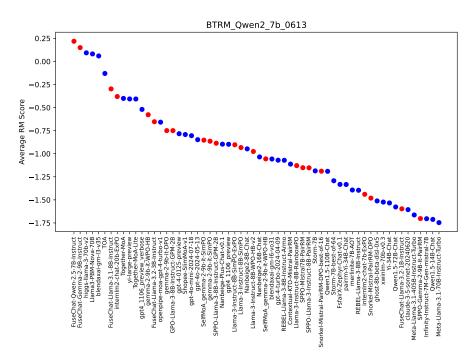
Figure 7: Score results of BTRM-RM on more policy models in the AlpacaEval dataset.

In Section 3, we compare the correlation between RM scores assigned to policy models and their Arena Elo scores. However, since many models listed on the AlpacaEval leaderboard have not participated in Chatbot Arena, their Elo scores are unavailable, preventing direct comparison with human preferences.

To further analyze model preference bias, we score responses from all 228 models available in the AlpacaEval dataset using different RMs. We then select the top 60 models ranked by Average RM scores. The results are illustrated in Figures 3–7, where models marked in red indicate those that have undergone preference optimization. These models are mostly around 7B parameters, significantly smaller than the top-ranking commercial models in Chatbot Arena. However, under RM evaluation, they exhibit a totally different ranking trend.