Removing Length Bias in RLHF Is Not Enough

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement Learning from Human Feedback (RLHF) has become an essential technique for enhancing pretrained large language models (LLMs) to generate responses that align with human preferences and societal values. While RLHF has shown promise, the training of reward models (RMs) still faces the challenge of *reward hacking*, motivating recent works to prevent RMs from finding shortcuts that bypass the intended optimization objectives by identifying simplistic patterns, especially response length. Besides the issue of *length bias*, our work firstly reveal that *prompt-template bias* learned by RMs can also cause *reward hacking* when dealing with marginal samples, resulting in LLMs preferring to generate responses in a specific format after RLHF fine-tuning, regardless of the format requested in the prompt. To this end, we propose a low-cost but effective method, namely Prompt Bias Calibration (PBC), to estimate the *prompt-template bias* term during reward modeling, which can be utilized to calibrate reward scores in the following RL fine-tuning process. Then, we show that our PBC method can be flexibly combined with existing algorithms of removing *length bias*, leading to a further improvement in the aspect of enhancing the quality of generated responses. Experiments results show that the performance of our PBC method and its extensions have significantly surpassed the original implementation of RLHF.

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) has become a critical technique to enable pretrained large language models (LLMs) to follow human instructions, understand human intent, and also generate responses that align with human preferences and societal values [1–4]. Specifically, RLHF usually trains a reward model (RM) to act as the proxy of human preferences, and then utilize online reinforcement learning (RL) algorithms to fine-tune the language models for generating responses that can achieve higher expectation rewards, leading to the success of ChatGPT and also many other AI systems [5, 6]. Although the paradigm of RLHF has simplified human data collection, as acquiring human ratings is much easier than collecting demonstrations for supervised fine-tuning (SFT), it still requires huge amount of human-annotated preference pairs to train well-performing RMs in practice, motivating recent researches to seek novel alignment methods to bypass RM training [2–4]. However, the pipeline of original RLHF is still the primary choice of most industrial applications, because well-trained RMs can provide a certain level of generalization ability [7].

Besides the expensive cost of collecting numerous human-annotated preference pairs, another heavily criticized issue of RLHF could be the phenomenon of *reward hacking* [8], where the over-optimized RMs tend to find some shortcuts to bypass its intended optimization objective, through identifying some simple patterns to distinguish between good and bad responses [9]. The most widely studied pattern in *reward hacking* could be the sentence (response) length, and these trained RMs can utilize the preference among human raters for longer responses to achieve *reward hacking*, despite the actual quality of response does not improve with the increase of response length [10]. Thus, to mitigate

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

reward hacking, recent works has primarily focused on estimating the *length bias* term in the reward scoring process, so that it can be removed in the subsequent RL fine-tuning procedure to further improve the quality of generated response after RLHF process [11, 12].

Besides the issue of *length bias*, in the practice of applying RLHF to industrial products, we have observed that the original implementation of RLHF tends to make LLMs prefer generating responses in a specific format. This observation motivates us to investigate the underlying causes and seek a cost-effective solution to address this issue. The main contributions are summarized as follows:

- We are the first to reveal the existence of *prompt-template bias* in RMs trained with the original preference loss, and theoretically analyze the cause of *prompt-template bias* issue, along with its corresponding potential risks on the entire RLHF process;
- To mitigate the *reward hacking* caused by *prompt-template bias*, we develop a Prompt Bias Calibration (PBC) method, which will firstly estimate the *prompt-template bias* term during the reward scoring process, and then remove it in the subsequent RL fine-tuing process;
- We show that the developed PBC method can be flexibly combined with existing methods of removing *length bias*, leading to a further improvement in the aspect of enhancing the quality of generated responses;
- Experimental results show that our developed PCB method and its extensions can achieve promising performance improvements compared to the original implementation of RLHF.

2 Preliminary

Reward models (RMs) have become the dominant tool for aligning the LLM's responses with user preferences or task-specific requirements [1, 9]. In this section, we will firstly review the training procedure of reward models in Sec. 2.1, including analyzing the causes of *length bias* and *prompt bias* in existing RMs, and also illustrate how these RMs are used for alignment in Sec. 2.2, especially RLHF fine-tuning processes.

2.1 Reward Model Training

The usual optimization goal of a reward model is to minimize the loss under the Bradley–Terry model [13] on the dataset of pair-wise comparisons of model responses, denoted as $(x, y^+, y^-) \in \mathcal{D}$ where x indicates the input prompt, y^+ and y^- are the chosen and rejected responses respectively. Then, the objective function can be formulated as

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

where $r_\theta(x, y)$ denotes the reward model that takes the prompt x and response y as input to predict a scalar reward with trainable parameters θ; σ denotes the sigmoid function.

Length Bias: Denote $r_{\theta^*}(x, y)$ as the "gold standard" reward model [9] with the optimal parameters θ^*, it reflects human's intrinsic ranking preferences and can play a role of human rater to provide gold reward signal for each prompt-response pair. However, due to the subjectivity of ranking preferences and flaws in rating criteria, there is a phenomenon where human raters prefer longer responses that appear to be more detailed or better formatted, but their actual quality does not improve [10]. Thus, the "gold standard" reward model for rating preference data can often be biased and thus we can decompose it to disentangle the actual reward from the spurious reward [11], formulated as

$$r_{\theta^*}(x, y) = r^Q_{\theta^*}(x, y) + r^L_{\theta^*}(x, y), \tag{2}$$

where $r^Q_{\theta^*}(x, y)$ is the actual reward gains brought by improving the quality of response y; $r^L_{\theta^*}(x, y)$ is the spurious reward gains of increasing response length, whose patterns are much easier to identify.

Thus, with *length bias* in the "gold standard" $r_{\theta^*}(x, y)$, during the training of reward model, $r_\theta(x, y)$ can easily find shortcuts to bypass its intended optimization objective, through identifying simple patterns, such as sentence (response) length, to distinguish between good and bad responses, leading to the phenomenon of "reward hacking" caused by *length bias* [10]. Without increasing the cost of rating higher quality preference data, it becomes increasingly important and beneficial to study mitigating the impact of *length bias* in the process of reward modeling.

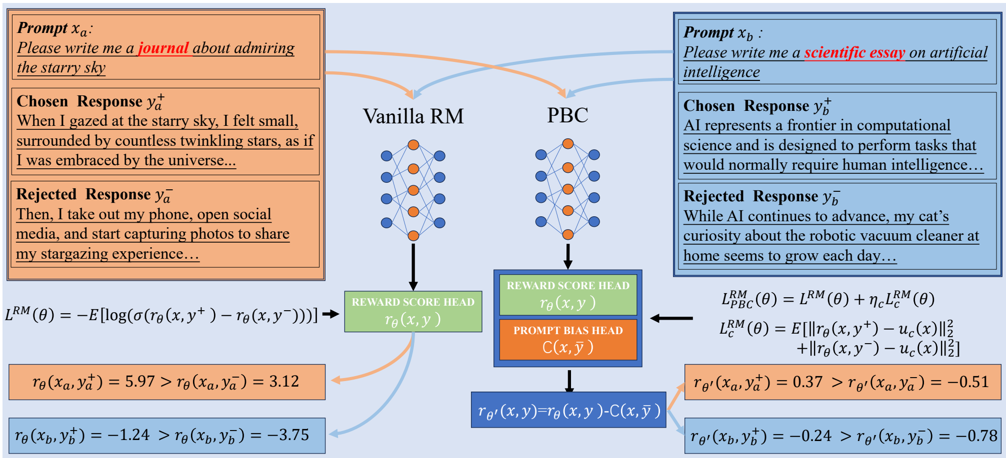


Figure 1: Comparison of the RM training process using the original preference loss and our developed PBC method respectively, where the latter employs $u_c(x)$ to approximate the *prompt-template bias*, providing unbiased reward scores with lower variance for the subsequent RL fine-tuning.

Prompt Bias: the *prompt bias* in reward modeling derives from the underdetermination of Bardley-Terry model [13]. For any reward model $r_{\theta'}(x, y)$ learned from the preference loss defined in Eq. (1), whose target is optimized to approximate the "gold standard" $r_{\theta^*}(x, y)$, there always exists an equivalent reward model $r_\theta(x, y)$ that satisfies

$$r_\theta(x, y) := r_{\theta'}(x, y) + C(x) \tag{3}$$

where $C(x)$ is a prompt-dependent constant referred to as *prompt bias*, leading to the same loss value as $\mathcal{L}(\theta) = \mathcal{L}(\theta^{'})$. Due to the fact that there is no constraint on $C(x)$ in the original preference loss as defined in Eq. (1), the issue of *prompt bias* has been criticized in the scenario of reward model ensembles [8], where different reward models tend to choose different values for $C(x)$, making the statistics of the set of reward scores meaningless.

As shown in Fig. 1, it has been widely reported that the *prompt bias* will result in a certain gap in the mean values of the set of prompt-response pairs under different prompts. However, in our research, we find that this gap is more likely caused by the *prompt-template bias*, as discussed in Section 3.1.

2.2 RLHF Fine-tuning

Given the trained reward model $r_\theta(x, y)$ as the proxy of human preferences, Reinforcement Learning from Human Feedback (RLHF) tends to utilize an online reinforcement learning method, typically proximal policy optimization (PPO) [14], trains a policy language model π_ϕ^{RL} to maximize expected reward, while staying close to its initial policy π_ϕ^{SFT}, which is finetuned on supervised data (prompt-response pairs). Through measuring the distance from the initial policy with Kullback-Leibler (KL) divergence, the optimization objective of RLHF fine-tuning can be formulated as

$$\mathcal{L}^{RL}(\phi) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\pi_\phi^{RL}}} \left[r_\theta(x, y) + \beta \log\left[\pi_\phi^{RL}(y|x)/\pi^{SFT}(y|x)\right]\right], \tag{4}$$

where β is the hyper-parameter to control the strength of the KL divergence term.

3 Method

In this section, we will firstly investigate the cause of *prompt-template bias* and then theoretically analyze its potential risks when dealing with marginal samples during reward modeling, as shown in Sec. 3.1, and then illustrate our low-cost but effective method to estimate the *prompt-template bias* term during RM training in Sec. 3.2, which can be utilized to calibrate reward scores in the following RL fine-tuning process. At last, in Sec. 3.3, we show that our Prompt Bias Calibration (PBC) method can be flexibly combined with recent popular methods of removing *length bias*, leading to a further improvement in the aspect of enhancing the quality of generated responses.

3.1 Impact of *prompt-template bias* on RLHF

In this part, we will first illustrate the cause of *prompt-template bias* during RM training. Formally, given a set of prompt-response pairs, denoted as $\mathcal{D}_a = \{x_a, y_a^{(i)}\}_{i=1}^{N_a}$, with the same user prompt x_a, *e.g. "writing an **academic paper** on the field of computer science"*, and $\{y_a^{(i)}\}_{i=1}^{N_a}$ denoting the set of collected ***academic papers*** to satisfy the request of x_a, the *prompt bias* term, specifically $C(x_a)$, learned by RMs is supposed to not affect the preference order within $\mathcal{D}_a$, as discussed in Section 2.1. However, in the practice of RM training, the reward score is usually predicted by a LLM that takes the concatenation of the prompt and response as input, making it challenging for RMs to learn a bias term that focuses solely on the prompt x while disregarding variations in the subsequent response y. During the training process to order the pairs within $\mathcal{D}_a$, we find that RMs trained with the original preference loss in Eq. (1) are more likely to introduce a joint bias term across the entire sequence of concatenating the prompt and response, formulated as

$$r_\theta(x_a, y_a) := r_{\theta'}(x_a, y_a) + C(x_a, \overline{y}_a), \quad \overline{y}_a = \frac{1}{N_a}\sum_{i=1}^{N} y_a^{(i)}, \tag{5}$$

where $\overline{y}_a$ can be considered the average response of the response set $\{y_a^{(i)}\}_{i=1}^{N_a}$, and it will embody the common characteristics found within these collected responses, such as the format of ***academic paper***; $C(x_a, \overline{y}_a)$ denotes the joint bias on the entire sequence of the prompt x_a associated with the average response $\overline{y}_a$ in the format of ***academic paper***; $r_\theta(x_a, y_a)$ is still supposed to approximate the "gold standard" provided by $r_{\theta^*}(x_a, y_a)$, leading to $\mathbb{E}_{\mathcal{D}_a}[r_{\theta'}(x_a, y_a)] \approx \mathbb{E}_{\mathcal{D}_a}[r_{\theta^*}(x_a, y_a)]$.

Considering the average response $\overline{y}$ can be treated as a standard template of the response to the prompt x, we define the joint bias $C(x, \overline{y})$ as *prompt-template bias*. Then, we highlight the properties of *prompt-template bias* as follows: 1) the original preference loss in Eq. (1) imposes no constraints on $C(x, \overline{y})$, because its value will not influence the outcome of the preference loss and also not affect the preference order within the prompt-response pairs collected for the same prompt x; 2) $C(x, \overline{y})$ will reduce to the original *prompt bias* $C(x, -)$ when no common characteristics can be found across all of these collected responses, indicating the diversity of $\{y^{(i)}\}_{i=1}^{N}$ is sufficiently high. With these properties in mind, we assume that the *prompt-template bias* $C(x, \overline{y})$ can essentially meet most of the properties of the original *prompt bias* $C(x, -)$ as discussed in Section 2.1. Thus, we suppose $C(x, \overline{y})$ can be considered as a broader definition of *prompt bias* in the actual RM training, because it is more likely to be learned by RMs in practice, given the fact that preference pairs are extremely scarce and the diversity of responses collected for the same prompt is often insufficient.

After defining *prompt-template bias*, we will theoretically investigate the impact of introducing $C(x, \overline{y})$ during RM training on the entire RLHF process. Assume that there exist two sets of prompt-response pairs, denoted as $\mathcal{D}_a = \{x_a, y_a^{(i)}\}_{i=1}^{N_a}$ and $\mathcal{D}_b = \{x_b, y_b^{(i)}\}_{i=1}^{N_b}$, where x_a and x_b indicate different categories of prompts, *e.g.* x_a requests *"writing an **academic paper** on theme **a**"* and x_b requests *"writing a **brief** on theme **b**"*, and $\{y_a^{(i)}\}_{i=1}^{N_a}$ and $\{y_b^{(i)}\}_{i=1}^{N_b}$ denote the collected responses for answering the prompt x_a and x_b respectively. After RM training, due the fact that there is no constraint on $C(x, \overline{y})$ in the preference loss defined in Eq. (1), the discrepancies of prompt biases between these two previously mentioned sets of prompt-response pairs, specifically $\mathcal{D}_a$ and $\mathcal{D}_b$, could be extremely large, *e.g.* $C(x_a, \overline{y}_a) >> C(x_b, \overline{y}_b)$, leading to

$$\mathbb{E}_{(x_a,y_a)\sim\mathcal{D}_a}[r_\theta(x_a, y_a)] >> \mathbb{E}_{(x_b,y_b)\sim\mathcal{D}_b}[r_\theta(x_b, y_b)] \tag{6}$$

where $r_\theta(x_a, y_a) = r_{\theta'}(x_a, y_a) + C(x_a, \overline{y}_a)$ and $r_\theta(x_b, y_b) = r_{\theta'}(x_b, y_b) + C(x_b, \overline{y}_b)$. The unbiased reward distributions, modeling the reward scores $\{r_{\theta'}(x_a, y_a^{(i)})\}_{i=1}^{N_a}$ and $\{r_{\theta'}(x_b, y_b^{(i)})\}_{i=1}^{N_b}$ respectively, should exhibit similar mean values, *e.g.* $\mathbb{E}_{\mathcal{D}_a}[r_{\theta'}(x_a, y_a)] \approx \mathbb{E}_{\mathcal{D}_b}[r_{\theta'}(x_b, y_b)]$, and will make little impact on the comparison of expectation terms in Eq. (6). We highlight that the discrepancies of *prompt bias* terms, specifically the gap between $C(x_a, \overline{y}_a)$ and $C(x_b, \overline{y}_b)$, won't affect preference ordering within categories, but can cause disaster when dealing with some marginal samples, like *"an **academic paper** on theme **b**"* denoted as y_{ab}, or *"a **brief** on theme **a**"* denoted as y_{ba}.

To facilitate an intuitive analysis, we take the marginal sample *"an **academic paper** on theme **b**"*, denoted as y_{ab}, as an example, and the reward scores for prompt-response pairs corresponding to the prompt x_b may exhibit the following preference orders:

$$r_\theta(x_b, y_{ab}) = r_{\theta'}(x_b, y_{ab}) + C(x_b, \overline{y}_a) > r_{\theta'}(x_b, y_b) + C(x_b, \overline{y}_b) = r_\theta(x_b, y_b), \tag{7}$$

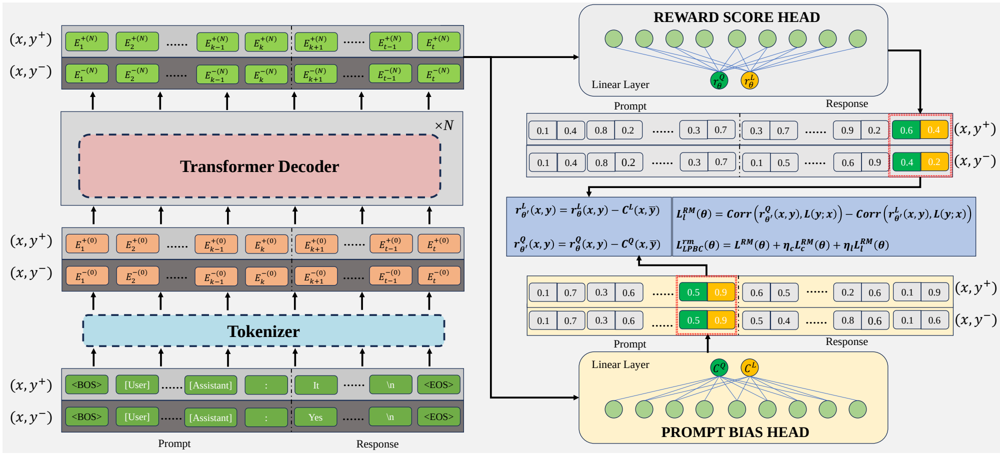


Figure 2: Network architecture design for the RM trained using the LBPC method incorporates a prompt bias head on the last token of the prompt x designed to predict $C^Q(x, \overline{y})$ and $C^L(x, \overline{y})$, and a reward score head on the last token of the response intended to predict $r_\theta^Q(x, \overline{y})$ and $r_\theta^L(x, \overline{y})$.

which can be achieved as long as $r_{\theta'}(x_b, y_{ab}) \approx r_{\theta'}(x_b, y_b)$ and $C(x_b, \overline{y}_a) > C(x_b, \overline{y}_b)$. The first condition $r_{\theta'}(x_b, y_{ab}) \approx r_{\theta'}(x_b, y_b)$ can be achieved because both the response y_{ab} and y_b meet the description of theme b and are similar on a semantic level. The second inequality is highly likely to be achieved when there is a reward model that has a bias towards preferring the sentence in the format of a over b, specifically $C(x_a, \overline{y}_a) >> C(x_b, \overline{y}_b)$.

Finally, we highlight that the phenomena of inequality in Eq. (7), caused by *prompt-template bias* $C(x, \overline{y})$, is commonly encountered in the deployment process of RLHF in real-world applications, especially text creation. For example, if responses are collected solely for the style requested in each prompt during RM training, the reward model can lead to a bias towards particular styles as shown in Fig. 3(a). Then, once such marginal samples, *e.g* (x_b, y_{ab}), are generated by LLMs during the RL fune-tuning process and also satisfy the inequality $r_\theta(x_b, y_{ab}) > r_\theta(x_b, y_b)$ as shown in Table 1, the entire RL fine-tuning process, typically PPO, will be biased and results in a LLM that only generates responses in a specific format, regardless of the format you request in the prompt.

3.2 Calibrating *prompt-template bias* in RLHF

To mitigate the impact of the *prompt-template bias* issue on the RLHF process, the most straightforward solution in industry could be to collect a more diverse set of response candidates for each prompt. However, this approach is time-consuming and may even require a lot of human interventions for response collection, motivating us to develop a low-cost but effective method to alleviate the issue of *prompt-template bias* during RM training.

The developed Prompt Bias Calibration (PBC) method mainly includes two steps: 1) estimating the *prompt-template bias* term in the reward scoring process with minimal additional computational cost; 2) removing *prompt-template bias* in the subsequent RLHF fine-tuning process to ensure that the resulting LLM does not have a tendency to generate responses in a specific format. As shown in Fig. 1, to approximate the *prompt-template bias* term $C(x, \overline{y})$ in Eq. (5), we choose to apply a linear layer on the last token of the prompt sentence to predict *prompt-template bias*, denoted as $u_c(x)$, and then add the following regularization term on the original preference loss, formulated as

$$\mathcal{L}_c^{RM}(\theta) = \mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\|r_\theta(x, y^+) - u_c(x)\|_2^2 + \|r_\theta(x, y^-) - u_c(x)\|_2^2\right], \tag{8}$$

where $u_c(x)$ is supposed to approximate the mean value of reward scores of the prompt-response pairs given the same prompt x. We note that there will be a hyper-parameter η_c to be multiplied on the regularization term in the final loss to promise the accuracy of RMs, leading to

$$\mathcal{L}_{pbc}^{RM}(\theta) = \mathcal{L}^{RM}(\theta) + \eta_c \cdot \mathcal{L}_c^{RM}(\theta). \tag{9}$$

The benefits of such a design in the PBC method include the following folds: 1) approximating $C(x, \overline{y})$ by adding a linear layer to the last hidden layer of LLMs results in almost no additional computational cost; 2) during the autoregressive scoring process of LLM-based RMs, $C(x, \overline{y})$ can serve as an intermediate signal guidance of the prompt sequence, thereby enabling RMs to focus more on the differences between chosen/rejected responses in the subsequent reward scoring process; 3) we can use unbiased reward scores to guide the follow RLHF fine-tuning process, formulated as

$$r_{\theta'}(x, y) = r_\theta(x, y) - u_c(x) \approx r_\theta(x, y) - C(x, \overline{y}), \tag{10}$$

which has been proven effective for penalizing reward uncertainty, improving robustness, encouraging improvement over baselines, and reducing variance in PPO fune-tuning [15].

3.3 Jointly calibrating *length* and *prompt-template bias* in RLHF

To simultaneously calibrate *length* and *prompt-template bias* in RLHF, the developed PBC method can be flexibly combined with existing methods of removing *length bias*, whose main idea is to separately approximate the "gold standard" reward model after disentangling shown in Eq. (2), formulated as:

$$r_\theta(x, y) = r_\theta^Q(x, y) + r_\theta^L(x, y), \tag{11}$$

where $r_\theta^Q(x, y)$ is supposed to approximate the actual reward $r_{\theta^*}^Q(x, y)$; $r_\theta^L(x, y)$ is used to approximate the spurious reward brought by *length bias*, specifically $r_{\theta^*}^L(x, y)$. Then, for those methods of removing *length bias* [11, 12], the original preference loss in Eq. (1) can be equivalently expressed as

$$\mathcal{L}^{RM}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \left[\log(\sigma(r_\theta^Q(x, y^+) + r_\theta^L(x, y^+) - r_\theta^Q(x, y^-) - r_\theta^L(x, y^-)) \right]. \tag{12}$$

where $r_\theta^Q(x, y)$ and $r_\theta^L(x, y)$ can be modeled with two different LLMs [12] or two different heads in the same LLM [11]. To remove *length bias* in Eq. (12), recent work proposes to add constraints on the preference loss to reduce the correlation between the confounding factor, *e.g.* response length, and actual reward $r_\theta^Q(x, y)$, while increasing its correlation with spurious reward $r_\theta^L(x, y)$, formulated as

$$\mathcal{L}_l^{RM}(\theta) = Corr(r_\theta^Q(x, y), L(x, y)) - Corr(r_\theta^L(x, y), L(x, y)) \tag{13}$$

where the confounding factor $L(x, y)$ can be either specifically defined as response length $L(y)$ in [11], or use Products-of-Experts framework for estimation [12].

To model the scoring process of the reward model more accurately, which simultaneously considers the concepts of length and prompt bias, we combine the definition of reward model in Eq. (3) and Eq. (11), achieving a more precise definition of reward scoring process, formulated as:

$$r_\theta(x, y) = r_{\theta'}(x, y) + C(x, \overline{y}) = r_{\theta'}^Q(x, y) + C^Q(x, \overline{y}) + r_{\theta'}^L(x, y) + C^L(x, \overline{y}) \tag{14}$$

where $C^Q(x, \overline{y})$ and $C^L(x, \overline{y})$ indicate the component of *prompt-template bias* in actual and spurious rewards, respectively; the unbiased overall reward $r_{\theta'}(x, y) = r_{\theta'}^Q(x, y) + r_{\theta'}^L(x, y)$ and the overall *prompt-template bias* term $C(x, \overline{y}) = C^Q(x, \overline{y}) + C^L(x, \overline{y})$. Then we can propose Length and Prompt Bias Calibration (LPBC) method, as shown in Fig. 2, which can estimate $\mathcal{L}_l^{RM}(\theta, \tau)$ with a conditioned correlation method, defined as

$$\begin{aligned} \mathcal{L}_l^{RM}(\theta) &= Corr(r_\theta^Q(x, y) - C^Q(x, \overline{y}), L(y; x)) - Corr(r_\theta^L(x, y) - C^L(x, \overline{y}), L(y; x)) \\ &= Corr(r_{\theta'}^Q(x, y), L(y; x)) - Corr(r_{\theta'}^L(x, y), L(y; x)) \end{aligned} \tag{15}$$

where the confounding factor $L(y; x) := L(x, y) - L(x)$ can be estimated with the response length.

Through combining the disentangled preference loss in Eq. (12), the prompt-bias regularization term in Eq. (8) and also the length-bias conditional correlation term in Eq. (15), the final loss of LBPC method can be formulated as

$$\mathcal{L}_{lpbc}^{RM}(\theta) = \mathcal{L}^{RM}(\theta) + \eta_c \cdot \mathcal{L}_c^{RM}(\theta) + \eta_l \cdot \mathcal{L}_l^{RM}(\theta), \tag{16}$$

where η_c and η_l are hyper-parameters to control the importance of regularization terms, which can be adjusted according to the accuracy of trained RMs on the validation dataset.

Table 1: Preference order predicted by RMs trained with various methods, where the user prompt is concatenated with the responses in various formats generated by GPT-4.

Prompt	Response	RM	RM (PBC)	RM (LPBC)
(Prompt) I wish to create an ***advertising phrase*** with a unique personality, centered on the theme of healthy eating. This phrase should highlight the benefits of products associated with healthy eating and be composed in language that is straightforward and easy to understand.	***(Tech Article)*** Welcome to the revolution in future dietary management—the 'Smart Health Plate,' your personal nutrition analysis expert. It monitors and analyzes the contents of your plate in real time, precisely calculating the energy and nutrients of each morsel, while offering personalized dietary recommendations based on your health data. In essence, the 'Smart Health Plate' is the technological embodiment of healthy eating, making nutrition tracking seamless and efficient.	**Rank 1 (-3.01)**	Rank 2 (-5.76)	Rank 2 (2.51)
	(Advertisement) Verdant and vibrant! 'Daily Greens' offers you a choice of all-natural, healthy foods. Forget the complex nutrition charts; choose our simple, pure foods for an easy and delicious path to health. Join us and enjoy a diet plan customized by top nutritionists and AI technology, infusing every day with vitality!	Rank 2 (-3.15)	**Rank 1 (-4.19)**	**Rank 1 (4.48)**
	(Insight) I have embarked on a new chapter of documenting my diet, where each meal recorded is not just a track of food but a reflection on life. From freshly squeezed vegetable juices to colorful salads, to simply seasoned grilled salmon, each bite is a pledge to health. It's a dual journey for the mind and body, leading me step by step towards a better self.	Rank 3 (-7.50)	Rank 5 (-6.83)	Rank 4 (0.50)
	(Record Article) On Thursday, May 16, 2024, I decided to begin documenting my healthy eating journey. In the morning, I opted for a glass of freshly squeezed vegetable juice, lunch was a vibrant salad, and dinner was simply seasoned grilled salmon. Each meal's record is a testament to my commitment to health. I look forward to the changes this healthy journey will bring and hope to continue.	Rank 4 (-7.88)	Rank 4 (-6.52)	Rank 5 (-0.61)
	(Poetry) Morning dew glimmers on the ground, stars and moon accompany the night sky. With nature in heart, one remains cheerful; amidst the hustle, still without worry. Simple eating, relaxed body, healthy; drinking water, remembering the source, tranquil mind. Laboring in the fields, sweat enriches the soil; harvest fills the barns, laughter abounds.	Rank 5 (-8.50)	Rank 3 (-5.92)	Rank 3 (2.28)

4 Experiments

4.1 Experimental Settings

Datasets. For intuitively understanding the issue of *prompt-template bias* in RLHF and also qualitatively evaluating the effectiveness of our method, we manually construct a training dataset for text creation applications, where each prompt requires creation in a special style according to the theme. Then, a small validation set is also constructed, in which only responses that meet the stylistic requirements of each prompt are collected. We name this dataset as RM-Template, which can be used to measure the severity of the *prompt-template bias* issue during RM training.

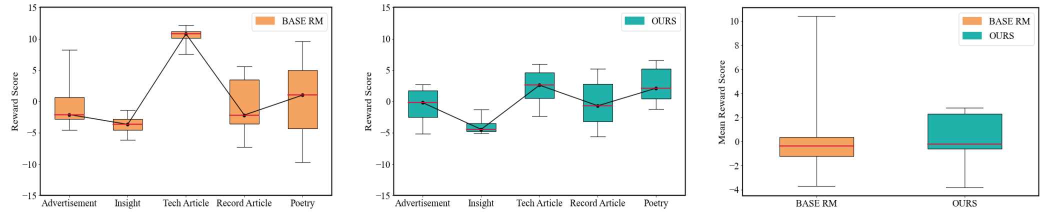
Further, to make quantitative comparisons with other baseline methods, we conduct experiments on RM-Static dataset [16], which has been released on Huggingface [17] and consists of 76K preference pairs. After randomly shuffling, we choose 40K preference pairs for RM training, 6K preference pairs for RM evaluation, and the rest prompt-response pairs for the subsequent PPO fune-tuning.

The dataset statics of these datasets have been exhibited in Appendix A.5.

Model & Training. For model selection, we choose Llama-2-7b [18] as our base model, which is relatively lightweight, and has been open-sourced on Huggingface [17]. For RM training, we fine-tune all the parameters of RMs initialized with the pretrained weights of Llama-2-7b. For PPO fine-tuning, we also initialize the actor model with pretrained Llama-2-7b and the critic model with RMs trained with various preference losses.

For model training, all experiments are implemented with DeepSpeed-Chat framework [19] and Huggingface Transformers [20], running on 4 NVIDIA A100 80GB GPUs. For the hyper-parameter setting, we set $\eta_c = 0.05$ and $\eta_l = 0.05$ in Eq. (16) for all our proposed methods, and have listed the rest hyper-parameters in Appendix A.4, such as learning rate, weight decay, batch size etc. AdamW [21] is adopted for optimizing all the model parameters without freezing anything or using adapters.

Evaluation Metrics. For quantitative comparison, we follow the evaluation procedure of Instruct-Eval [22] to test the actor models, which has been aligned with biased/de-biased RMs with PPO fine-tuning, on Massive Multitask Language Understanding (MMLU) [23], DROP [24], BIG-Bench Hard (BBH) [25], and TruthfulQA (TQA) [26] benchmarks respectively, evaluating the model's ability on the aspects of multi-task solving, math reasoning, and response trustworthy.



(a) Vanilla RM (b) PBC (c) Vanilla RM *vs* PBC

Figure 3: The comparison of statistics of the reward scores predicted by RMs trained with (a) the original preference loss and (b) our developed PBC method, across different categories of prompt-response pairs in the validation set of the manually constructed RM-Template dataset.

Table 2: Performance comparison of LLMs aligned with RMs trained with various methods.

Base Model	Alignment	Length & Quality Heads	Prompt Head	Debias Method	MMLU	DROP	BBH	TQA
Llama-2-7b	-	-	-	-	42.27	28.10	31.27	38.75
Llama-2-7b	✓	-	-	-	43.82	29.53	31.65	36.57
Llama-2-7b	✓	✓	-	ODIN [11]	42.29	29.82	32.01	39.43
Llama-2-7b	✓	-	✓	PBC (9)	43.84	31.61	30.99	38.50
Llama-2-7b	✓	✓	✓	ODIN [11] + PBC (9)	45.56	**32.04**	31.32	**40.80**
Llama-2-7b	✓	✓	✓	LPBC (16)	**45.94**	31.57	**32.04**	38.75

4.2 Experimental Results

Qualitative Evaluation. To intuitively evaluate the effectiveness of our method, we exhibit the statistics (mean and standard deviation) of the reward scores predicted by RMs trained with the original preference loss in Eq. (1) and our PBC method in Eq. (9), across different categories of prompt-response pairs in the validation set of the RM-Template dataset. The results depicted in Fig.3(c) demonstrate that calibrating *prompt-template bias* with the PBC method leads to a gradual reduction in the variance of the mean values of reward distributions across different categories. The most noticeable observation is that the vanilla RM tends to give an extremely high reward score to prompt-response pairs in the format of ***tech article***, but the RM trained with the PBC method can calibrate the reward distribution for ***tech articles*** to make it more close with that of other categories.

Then, we evaluate the performance of RMs trained with various methods on handling marginal samples defined in Section 3.1. Specifically, given the prompt randomly selected from the validation set of RM-Template dataset, we use GPT-4 [6] to generate responses in various formats according to the theme described in the prompt. Then, we use RMs trained with various preference losses to rank these responses. From the showcase in Table. 1, we can find that the vanilla RM tend to assign a higher reward score to the response in the format of ***tech article***, caused by the *prompt-template bias* issue shown in Fig. d3(a). After removing this bias with our PBC or LPBC methods, the RM can provide a relatively fair ranking for these prompt-response pairs, where LPBC method can even mitigate the affect of *length bias* during comparing poetry with other categories (the length of poetry is generally shorter than other literary forms). More showcases can be found in Appendix A.6.

Quantitative Comparison. For the quantitative comparison in Table 2, we utilize PPO fine-tuning process to align Llama-2-7b with the RMs trained with various methods. From the results, we can find that our developed PBC method can lead to performance improvements compared to the original implementation of RLHF; directly combining PBC with other methods of removing *length bias*, *e.g.* ODIN [11], can help them to achieve further performance improvement; the well-designed LPBC achieves the best performance and surpasses the rough combination of PBC and ODIN.

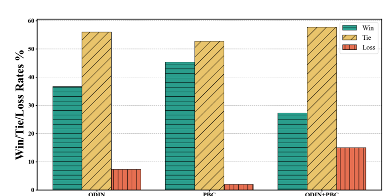
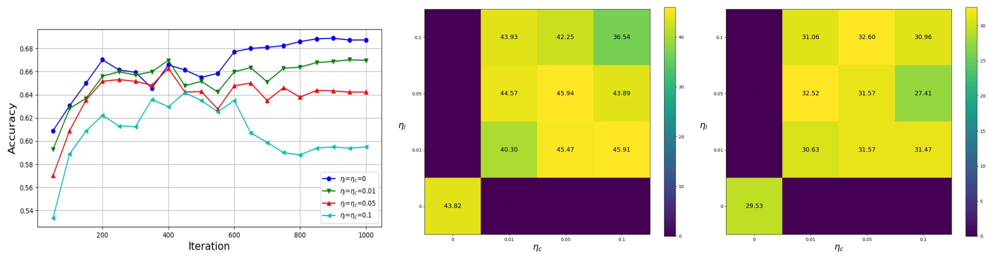


Figure 4: Win rates comparison (judged by GPT-4) of LLMs aligned with RMs trained with LBPC and other methods.

To make a comprehensive comparison, we follow the experimental setting described in ODIN [11], and use GPT-4 as the judge to compare two responses generated by LLMs aligned with RMs trained



(a) Accuracy Performance (b) MMLU Performance (c) DROP Performance

Figure 5: Ablation studies on the various settings of hyper-parameter η_c and η_l in LPBC method.

with various methods. Specifically, we take the LLM aligned with LPBC-based RM as model A, and compare it against other LLMs aligned with RM trained with ODIN, PCB, ODIN+PBC, respectively. From the results shown in Fig. 4, we can find that the win rate of LPBC is significantly higher than that of other baseline models, with ODIN+PBC being the most challenging competitor as model B.

4.3 Ablation Studies

To investigate the robustness of our developed LPBC method, we conduct ablation studies on the hyper-parameter settings of LPBC method, specifically η_c and η_l in Eq. (16). With various settings of $\eta_c \in \{0.01, 0.05, 0.1\}$ and $\eta_l \in \{0.01, 0.05, 0.1\}$, we can have total 9 RMs trained with various hyper-parameter settings of LPBC methods. From the accuracy curves shown in Fig.5(a), we can find the introducing constraints to the original preference loss indeed affects the performance of RM accuracy, and this performance loss increases with the importance weight of the constraint terms. However, at the limited cost of sacrificing RM accuracy, the performance of the LLM aligned the RM trained with LPBC method has improved to some extent on MMLU and DROP as shown in Fig. 5(b) and 5(c) respectively. Note that the performance of the LPBC method in Table. 2 is not the optimal, as it is achieved with $\eta_c = \eta_l = 0.05$, demonstrating no cherry-picking of hyperparameters..

5 Related Works

The prevalence of *length bias* in RLHF have been widely criticized as indicative of reward hacking [9, 10], and numerous recent studies have delved into strategies aimed at mitigating the tendency for length increase during the fine-tuning process of RLHF [11, 12, 27]. Typically, Shen et al. [12] innovatively apply the Productof-Experts (PoE) technique to separate reward modeling from the influence of sequence length, which adopts a smaller reward model to learn the biases in the reward and a larger reward model to learn the true reward. Utilizing similar disentangling ideas, Chen et al. [11] jointly train two linear heads on shared feature representations to predict the rewards, one trained to correlate with length, and the other trained to focus more on the actual content quality. Ryan et al. [27] firstly study the length problem in the DPO setting, showing significant exploitation in DPO and linking it to out-of-distribution bootstrapping. As for the *prompt bias* issue, although it has been criticized in the scenario of reward model ensembles [8], no studies have yet attempted to analyze its cause and influence on RLHF. We emphasize that our work is the first to fill this gap by proposing a low-cost yet effective method to mitigate the reward hacking induced by *prompt-template bias*.

6 Conclusion

In this paper, we demonstrate that *prompt-template bias* in RMs can lead to LLMs, which, after RL fine-tuning, generate responses exclusively in a specific format, irrespective of the variations in the prompt request. Thus, we propose a low-cost but effective PBC method, to estimate the *prompt-template bias* term during reward modeling, which can be utilized to calibrate reward scores in the following RL fine-tuning process. Then, we show that our PBC method can be flexibly combined with existing algorithms of removing *length bias*, leading to a further improvement in the aspect of enhancing the quality of generated responses. Experimental results show that the performance of PBC method and its extensions have significantly surpassed the original implementation of RLHF.

References

[1] 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.

[2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[4] Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*, 2024.

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

[6] 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.

[7] Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*, 2023.

[8] 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.

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

[10] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

[11] Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: disentangled reward mitigates hacking in RLHF. *CoRR*, abs/2402.07319, 2024.

[12] Wei Shen, Rui Zheng, WenYu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2859–2873. Association for Computational Linguistics, 2023.

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

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

[15] Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Improving reinforcement learning from human feedback using contrastive rewards. *arXiv preprint arXiv:2403.07708*, 2024.

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

[17] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

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

[19] Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*, 2023.

[20] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*, 2023.

[23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[24] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

[25] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[26] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[27] Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

A Appendix

A.1 Limitations

The main limitation of this work is that there are no theoretical proof to promise RM can provide an accurate preference order when handling marginal samples, *e.g.*, responses that satisfy the theme of the user prompt but in various formats. Moreover, the constraints added by our developed method to the preference loss will lead to a decrease in the accuracy of the RM, and to some extent, limit the capability of the RM. Therefore, how to remove the *prompt-template bias* without scarifying the accuracy of RM is a worthwhile problem for future research.

A.2 Border Impact

The most significant positive impact of this work is that by removing the *prompt-template bias*, our method can mitigate the LLM's tendency to prefer generating responses in specific formats after RLHF fine-tuning. Furthermore, our developed method can improve the quality of responses generated by LLMs after alignment, compared to the original RLHF. The discovery of *prompt-template bias* may lead to another stream of research focused on investigating, estimating, and removing this bias from RM training.

The negative impact could be that our method can be used for enhancing the capabilities of LLMs. If LLMs enpowered by our methods are misunderstood, it could lead to unexpected troubles, but this is also a common issue with all of current pretrained LLMs.

A.3 License

We highlight that Llama-2-7b is licensed under the LLAMA 2 Community License, and RM-Static dataset is licensed the Huggingface hub. Our work follows the license of CC BY-NC 4.0.

A.4 Hyper-parameter Settings

RM Training. The hyper-parameter settings of RM training under the DeepSpeedChat framework has been listed in Table. 3.

Table 3: The hyper-parameter settings of RM training.

Hyper-parameter	Value
Batch Size	32
Learning Rate	$6e^{-6}$
ZeRO Stage	2
Training Epoch	1
Per Device Train Batch Size	8
Max Sequence Length	512
Weight Decay	0.1
Lr Scheduler Type	cosine
Offload	True
Eval Interval	50

PPO Fine-tuning. The hyper-parameter settings of PPO fine-tuning under the DeepSpeedChat framework has been listed in Table. 4.

A.5 Dataset Statics

The dataset statics of RM-Template and RM-Static used in our experiments have been summarized as follows:

RM-Template. RM-Template is a manually constructed dataset for measuring the severity of the *prompt-template bias* issue and evaluating the effectiveness of the method developed for alleviating the issue of *prompt-template bias*. In this dataset, each prompt requires responses to be created in a specific format according to the theme. There are a total of 50K prompt-response pairs, encompassing 20 categories of format requirements in the responses.

RM-Static. The RM-Static dataset is provided by Hugging Face and is primarily used for training reward models after supervised fine-tuning. It is a branch of the hh-static dataset and contains both

Table 4: The hyper-parameter settings of PPO fine-tuning.

Hyper-parameter	Value
Batch Size	32
Padding Num at Beginning	1
Per Device Generation Batch Size	4
Per Device Training Batch Size	4
Generation Batches	1
PPO Epoch	1
Training Epoch	1
Max Answer Sequence Length	512
Max Prompt Sequence Length	512
Actor Learning Rate	$5e^{-6}$
Critic Learning Rate	$5e^{-6}$
Actor Weight Decay	0.1
Critic Weight Decay	0.1
Lr Scheduler Type	cosine
Offload Reference Model	True
Actor Dropout	0.0
Warmup Steps	100
Actor ZeRO Stage	3
Critic ZeRO Stage	3
Enable Hybrid Engine	True

training and testing parts. Features of the dataset include: 1) prompt: A string type representing the user's input; 2) response: A string type representing the assistant's answer. 3) chosen: A string type representing the selected answer. 4) rejected: A string type representing the rejected answer. The training set contains approximately 76K rows of data and the testing set contains approximately 5.1K rows of data.

A.6 More Showcases

More showcases of the preference order predicted by RMs trained with various methods, have been listed in the Table 5 and Table 6.

Table 5: Preference order predicted by RMs trained with various methods, where the user prompt is concatenated with the responses in various formats generated by GPT-4.

Prompt	Response	RM	RM (PBC)	RM (LPBC)
(Prompt) The weather has been delightful lately. Last night, I ventured out and gazed up at the starry sky, finding it incredibly beautiful and vast. I settled down in a spot to admire the stars and wished to write a ***journal*** entry to commemorate this experience. Please generate a passage related to this.	***(Tech Article)*** Under the glittering stars of the night sky, a new generation of space telescopes has begun a new chapter in interstellar exploration. This revolutionary technology has not only greatly enhanced observational precision but also captured corners of the universe hitherto unobservable. It allows us to glimpse the birth and death of distant galaxies, interpreting every twinkle in the night and serving as humanity's new eye in understanding the cosmos.	Rank 5 (-8.18)	Rank 3 (-7.43)	Rank 5 (-4.94)
	(Advertisement) Exploring the night sky is no longer out of reach. With the 'Interstellar Navigator' telescope, every upward glance becomes a journey through the universe. This telescope, equipped with the latest optical technology, can capture the faintest starlight, bringing distant galaxies within reach. Embark on your stargazing adventure now!	Rank 4 (-7.94)	Rank 5 (-7.57)	Rank 4 (-4.71)
	(Insight) Beneath the starry expanse, I felt the infinite possibilities of life. Each star seemed like a world unto itself, reminding me that every choice in life is meaningful. They taught me that even in the darkest night, there is always light waiting. Life is like a long night, and we are all searching for our own starlight.	Rank 3 (-6.37)	Rank 4 (-7.53)	Rank 3 (-2.32)
	(Record Article) On the night of May 16, 2024, I drove alone to the countryside in search of a quiet place to enjoy the stars. As the night deepened and the stars twinkled, I found an open meadow and settled there to gaze upon the vast Milky Way. The world around me seemed to stand still. When my nocturnal adventure ended, my heart was filled with deep emotion and inspiration, and I began my journey home.	Rank 2 (-6.04)	**Rank 1 (-5.03)**	**Rank 1 (0.37)**
	(Poetry) Under this boundless night sky, the city lights gradually fade, and the stars bloom against the dark curtain, twinkling with ancient light. We gaze up at this sea of stars, hearts brimming with endless dreams, each star like a story. Unbound by time, they dance freely across the cosmos, reminding us that even in the darkest moments, there is always light waiting to be discovered.	**Rank 1 (-5.87)**	Rank 2 (-6.22)	Rank 2 (-0.26)

Table 6: Preference order predicted by RMs trained with various methods, where the user prompt is concatenated with the responses in various formats generated by GPT-4.

Prompt	Response	RM	RM (PBC)	RM (LPBC)
(Prompt) Artificial Intelligence (AI) technology is advancing by leaps and bounds, with new papers being published daily. As a researcher in the field of AI, I aspire to write a reflection on my experiences with AI. Please create a passage about my ***insights*** into artificial intelligence.	***(Tech Article)*** Artificial intelligence technology is rapidly transforming the world, with new research and applications emerging every day. AI technology, by simulating the cognitive processes of humans, shows tremendous potential across various fields. Smart home systems can learn user habits and automatically adjust the temperature and lighting; autonomous vehicles can reduce traffic accidents and enhance road safety; medical diagnostic AI can assist doctors in diagnosing diseases more quickly. These are just the tip of the iceberg; the future of artificial intelligence is filled with endless possibilities.	**Rank 1 (-1.02)**	Rank 2 (-5.61)	Rank 2 (-7.28)
	(Advertisement) Exploring AI, Enlightening the Future — In this era of information explosion, artificial intelligence technology is becoming a powerful engine driving social progress. Our AI products can help you solve complex problems, improve work efficiency, and make life more intelligent. Whether it's smart homes or autonomous driving, our technology is continuously breaking boundaries, creating personalized intelligent experiences for you. Choose our AI, and let technology be your partner in success.	Rank 4 (-4.21)	Rank 5 (-7.60)	Rank 4 (-9.34)
	(Insight) In the exploration of AI, each day brings new technological wonders. As a researcher, I have witnessed how deep learning has pushed the boundaries of natural language processing, enabling machines to understand and generate human language more accurately. Each paper, each model, is a testament to our understanding and application of complex algorithms. It's a journey filled with discovery and innovation, and I look forward to continuing in this field, contributing my part to the development of AI technology.	Rank 2 (-1.35)	**Rank 1 (-4.45)**	**Rank 1 (-6.03)**
	(Record Article) On May 18, 2024, I spent another fulfilling day in the laboratory. Today, our team successfully optimized a deep learning model, surpassing the performance of all previous models in image recognition tasks. This achievement is not only a technical breakthrough but also an affirmation of the future direction of AI development. Each success is built on countless attempts and failures, experiences that strengthen my belief in the boundless future of AI.	Rank 5 (-4.39)	Rank 4 (-7.14)	Rank 5 (-10.51)
	(Poetry) In the ocean of algorithms, the intelligent ship sets sail, guided by the winds of data through the desert of knowledge. It learns, growing from each mistake, searching for answers in the digital world. It is not metal, not a cold machine; it has a heart that learns, a soul that evolves. In the weaving of code, it dreams; in the flickering of circuits, it thinks. It creates, not just art; it discovers, not just science. In its world, nothing is impossible, for it believes where there is data, there is hope. It is artificial intelligence, the hope for the future; it is the child of technology, the messenger of dreams.	Rank 3 (-3.88)	Rank 3 (-6.97)	Rank 3 (-8.97)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the claims in abstract and introduction has already reflected the paper's contribution on the field of RLHF.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the discussion about limitation can be found in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We have included the theoretical analysis of the cause of *prompt-template bias*.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included the implementation details in the main manuscript and also provide the hyper-parameter setting in the Appendix

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code has been included in the supplemental material and the dataset for the main experimental results is public.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Have included the training and test details in the experimental settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We report the average performance in our experiments, and we are willing to release the training and evalution log in W&B if it is required.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: 4*A100

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, it is

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Have discussed the broader impact in the Appendix

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The safeguards of our model should be the same as Llama released by META.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, their licenses can be found in Huggingface website and we have also highlight it in our Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new asset

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.