Alleviating Shifted Distribution in Human Preference Alignment through Meta-Learning

Shihan Dou[1], Yan Liu[1], Enyu Zhou[1], Songyang Gao[1], Tianlong Li[1], Limao Xiong[1], Xin Zhao[2], Haoxiang Jia[3], Junjie Ye[1], Rui Zheng[1], Tao Gui[4*], Qi Zhang[1,5], Xuanjing Huang[1,6*]

[1]School of Computer Science, Fudan University, Shanghai, China
[2]Ant Group, Shanghai, China
[3]School of Computer Science, Peking University, Beijing, China
[4]Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China
[5]Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China
[6]Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Shanghai, China
shihandou@foxmail.com, tgui@fudan.edu.cn

Abstract

The capability of the reward model (RM) is crucial for the success of Reinforcement Learning from Human Feedback (RLHF) in aligning with human preferences. However, as training progresses, the output space distribution of the policy model shifts. The RM, initially trained on responses sampled from the output distribution of the early policy model, gradually loses its ability to distinguish between responses from the newly shifted distribution. This issue is further compounded when the RM, trained on a specific data distribution, struggles to generalize to examples outside of that distribution. These two issues can be united as a challenge posed by the shifted distribution of the environment. To surmount this challenge, we introduce MetaRM, a novel method leveraging meta-learning to adapt the RM to the shifted environment distribution. MetaRM optimizes the RM in an alternating way, by preserving both the preferences of the original preference pairs, as well as maximizing discrimination power over new examples of the shifted distribution. Extensive experiments demonstrate that MetaRM can iteratively enhance the performance of human preference alignment by improving the RM's capacity to identify subtle differences in samples of shifted distributions.

Introduction

Reinforcement learning from human feedback (RLHF) provides a pivotal technique to ensure that the behavior of AI systems aligns with the intentions of their designers and the expectations of users (Bai et al. 2022; Ouyang et al. 2022; Zheng et al. 2023b). RLHF is executed in two primary stages. The initial stage involves training a reward model using preference data, which is collected from a substantial number of crowdsource workers. The second stage entails the application of reinforcement learning (RL) to fine-tune the large language model (LLM), to maximize the reward. In this process, the reward model plays a pivotal role, as its performance significantly impacts the effectiveness of alignment with human preference (Eschmann 2021; Gao, Schulman, and Hilton 2022).

However, the reward model faces generalization challenges caused by the environment distribution shifts during the RL phase, as shown in Figure 1. Specifically, as the RL training progresses, the optimization of the policy model causes shifts in its output space distribution. Consequently, the reward model, initially trained on the preference pairs derived from the output distribution of the early policy model, gradually fails to distinguish between responses from the newly shifted distribution. This issue is also mirrored in out-of-distribution (OOD) scenarios. The reward model trained on data from a specific distribution struggles to identify subtle differences in OOD samples and has poor performance on such shifted distribution (Casper et al. 2023; Wulfe et al. 2022). Although researchers propose to iteratively annotate preference pairs and fine-tune the reward model to adapt it to the shifted environment (Touvron et al. 2023; DeepSeek-AI 2024), continuously collecting new data is resource and time-intensive. The approach of efficiently adapting the reward model to the shifted distribution remains insufficiently explored.

To solve this challenge, we introduce MetaRM, a novel approach that adapts the reward model to the shifted distribution and restores its distinguishing ability by using meta-learning. MetaRM utilizes an alternating optimization way to train the reward model by minimizing the loss on the original preference pairs, particularly those data that can maximize the discrimination power to responses of the target-shifted distribution. In this way, we can bridge the gap between the preference data distribution and the target-shifted distribution of the environment. It ensures that the reward model not only performs well on the original preference distribution but also can distinguish the differences in samples of the target-shifted distribution. In terms of implementation, MetaRM can constantly enhance the performance of human preference alignment by iteratively adapting the reward model to the output distribution of the new policy model, achieving an iterative RLHF. Additionally, MetaRM

*Corresponding authors
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

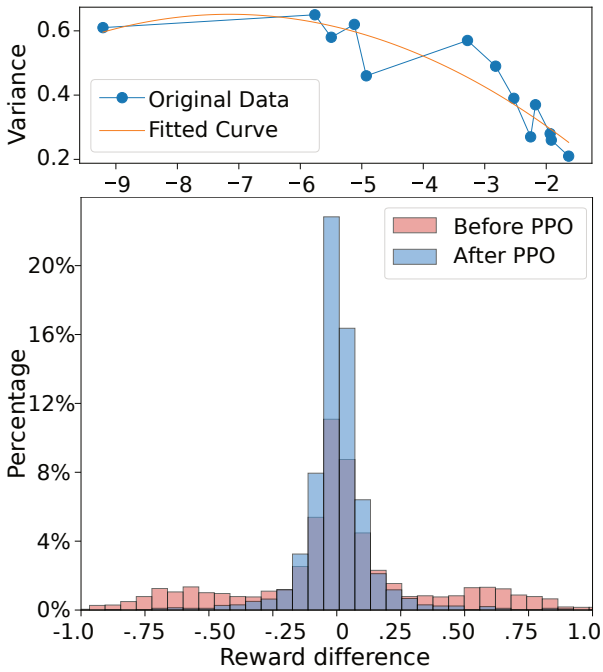

Figure 1: **(Upper)** Variance of reward difference distribution. We randomly select 1,000 prompts in the training set and then sample two responses for each prompt from the output distribution of the policy model and compute the difference between the rewards over time. As training progresses, the distribution of output space shifts, leading the RM to gradually fail to distinguish between responses. **(Bottom)** Reward Difference Distributions. We sample two responses from a specific distribution for each prompt and obtain the difference between the rewards. For **red** and **blue** plots, the responses are sampled from the output distribution of the initial policy model and the latest policy model trained after PPO, respectively. The RM provide close rewards for the different responses in most queries. These indicate that the RM fails to capture subtle differences between responses under conditions of shifting environment distribution.

also improves the capability of the reward model in OOD scenarios by training only on original preference pairs.

To evaluate the effectiveness of MetaRM, we conduct extensive experiments on the Anthropic's HH-RLHF (Bai et al. 2022) and OpenAI's summarization (Stiennon et al. 2020b) datasets. The experimental results demonstrate that MetaRM can constantly restore the reward model's distinguishing ability to those responses sampled from the shifted distribution by iteratively training it on original preference data, achieving improvement of the LLM in 3 to 4 rounds. In addition, we also evaluate MetaRM in an OOD setting. The experimental results reveal that it outperforms other reward modeling approaches in LLM alignment by enhancing the ability to discriminate subtle differences in OOD samples. The main contributions of our paper are as follows:

- We introduce MetaRM, a novel method that adapts the reward model to the shifted distribution through meta-learning, to enhance its ability to distinguish responses sampled from the shifted distribution.
- Extensive experiments show that MetaRM can iteratively improve the performance of LLM alignment by constantly training the reward model on the original preference pairs.
- MetaRM also enhances the ability of the reward model trained only on specific distribution preference data to effectively discriminate OOD samples, without the need for labeling pairs data on the target distribution.

Related Work

Reinforcement Learning from Human Feedback. Previous studies have demonstrated that RLHF (Bai et al. 2022; Ouyang et al. 2022) is a key component of training state-of-the-art LLMs, such as OpenAI's GPT-4 (OpenAI 2023) and Meta's Llama 2 (Touvron et al. 2023). Meanwhile, it also can improve various tasks, such as summarization (Stiennon et al. 2020a; Ziegler et al. 2019), dialogue (Bai et al. 2022), translation (Bahdanau et al. 2016), and make LLMs more helpful, honest, and harmless (3H) (Thoppilan et al. 2022; Ouyang et al. 2022). RLHF involves two main steps: first, using preference data collected from a large number of crowdsource workers to train a reward model. Secondly, using reinforcement learning methods to optimize the language model to maximize the reward. The reward model plays a crucial role in the RLHF process, so modeling a robust reward model is crucial for the RLHF (Ramé et al. 2024; Lee et al. 2023).

Distribution Shift in Reward Models. Researchers have attempted to obtain a robust reward model by accurately modelling human preferences to boost the ability of the reward model and improve the performance of LLMs (Coste et al. 2023; Shen et al. 2023; Pace et al. 2024). Although these approaches can model reward models somewhat better, they are still suffering from the distribution shift in the RL training phase (Casper et al. 2023; Pikus et al. 2023). Casper et al. (2023) illustrates that distribution shifts can decrease the credibility of the reward model. Additionally, Krueger, Maharaj, and Leike (2020) analyses that samples with overestimated rewards will become gradually more, which may lead to stagnation in the RL training process. Ramé et al. (2024) ensemble multiple reward models to mitigate the distribution shift and hence the reward overoptimization problem. Touvron et al. (2023) propose to iteratively collect preference pairs and fine-tune the reward model to adjust it to the new distribution. However, continuously collecting new data is resource and time-intensive. In contrast to these approaches, our method focuses on how to alleviate distribution shifts and align with out-of-distribution without labeling the data.

Meta-Learning. Meta-learning generally seeks to improve the models to adapt to new skills, unseen tasks, or new distributions (Finn, Abbeel, and Levine 2017; Li et al. 2019). With the advancement of LLMs, researchers have also introduced meta-learning into language models to enhance performance across various language-related tasks (Hospedales et al. 2021; Bansal et al. 2020; Min et al. 2021). Chen et al.

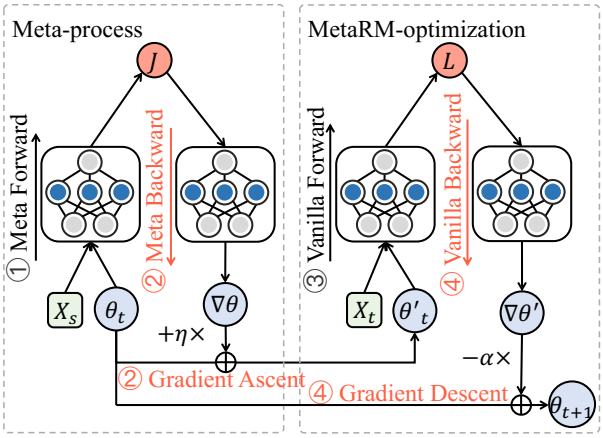

Figure 2: The optimization process of MetaRM. MetaRM contains four simple steps: 1. Compute the difference loss on responses sampled from the shifted distribution. 2. Calculate the gradient of this loss wrt. the RM parameters θ_t and adjust the parameters according to the ascent direction. 3. Compute the vanilla loss on the original preference pairs using the updated parameters $\theta_t^{'}$. 4. Calculate the gradient of the vanilla loss wrt. $\theta_t^{'}$ and optimize the original parameters θ following the descent direction.

(2021) introduce meta-learning into in-context learning in language models, focusing on enhancing the adaptability of these models to new tasks with limited data. Jia (2024) efforts to train the reward model in multi-tasks by using meta-learning, to improve the model's generalization ability. Dou, Yu, and Anastasopoulos (2019) explore meta-learning in low-resource natural language understanding tasks. Unlike these methods, our approach employs meta-learning to address distribution shift issues, enabling the reward model to distinguish out-of-distribution samples without the need for labeled data. Our proposed approach also can be utilized for iterative RLHF optimization.

Method

In this section, we elaborate on the methodological details of MetaRM, and provide a detailed explanation of the optimization objective of our method.

MetaRM

Our goal is that when the distribution of the environment shifts as the PPO training progresses, the reward model should still maintain the ability to distinguish new distribution responses, while modeling the human preference from original preference pairs. The key insight of MetaRM is that iteratively training the RM by minimizing the loss on the original preference pairs, particularly those pairs that can maximize the discrimination power over responses sampled from the shifted distribution. The optimization process of our proposed method MetaRM is shown in Figure 2.

The vanilla reward model is trained on a preference pairs dataset which contains comparisons between two responses under the same prompts (Bai et al. 2022; Ouyang et al. 2022). Formally, for a given prompt x inputted to the supervised fine-tuning (SFT) model $\pi^{\text{SFT}}(y|x)$, the two responses generated by π^{SFT} are denoted as y_1 and y_2. The labeller provides a preference for these two responses y_1 and y_2, denoted $y_w \succ y_l$, where y_w is the response more consistent with prompt x. Let the training dataset of the RM is $\mathcal{D} = \{(x^i, y_w^i, y_l^i), 1 \leq i \leq N\}$ and N is the number of preference pairs. The loss function of the vanilla reward model can be simplified as follows:

$$\mathcal{L}_\theta = -E_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))], \quad (1)$$

where r_θ denotes the reward model which is often initialized from the SFT model π^{SFT} and θ is the parameters of the reward model r_θ.

When putting reinforcement learning in the realm of large language models, the environment distribution and the output space distribution of the policy model $\pi^{\text{RL}}(y|x)$ are identical. It means that as $\pi^{\text{RL}}(y|x)$ is optimized, the environment distribution shifts. We find that the RM fails to effectively distinguish between responses sampled from the same prompt in the shifted environment, as shown in Figure 1. To measure the reward model's ability to distinguish the different responses under the same prompts, we define the difference loss function $\mathcal{J}_\theta$ of the reward model r_θ. Formally, let $s = \{s_i, 1 \leq i \leq k\}$ be the sequence of responses generated multiple times by the policy model $\pi^{\text{RL}}(y|x)$ under the same prompt x, where k denotes the number of responses. The difference function $\mathcal{J}_\theta$ can be written as follows:

$$\mathcal{J}_\theta = \frac{2}{k^2} \sum_{i=1}^{k} \sum_{j=i+1}^{k} \sigma(|r_\theta(x, s_i) - r_\theta(x, s_j)|). \quad (2)$$

It represents the degree of difference in the rewards given by r_θ for responses s. When the environment distribution shifts, $\mathcal{J}_\theta$ tends to have a lower value. In contrast, a reward model with a higher loss value indicates that it has a remarkable ability to differentiate subtle differences in responses.

Inspired by meta-learning, to restore the reward model's ability to distinguish responses sampled from a shifted distribution, we introduce an alternating optimization (*i.e.,*

Algorithm 1: The optimization process of MetaRM.

Require: $\theta, \mathcal{D}, \mathcal{S}, n, m$
Require: η, α
for $t = 0, \cdots, T-1$ **do**
  Sample a mini-batch $X_t = \{(x^i, y_w^i, y_l^i), 1 \leq i \leq n\}$ of size n from the preference pairs dataset $\mathcal{D}$
  Sample a mini-batch $X_s = \{(x^i, s^i), 1 \leq i \leq m\}$ of size m from the meta dataset $\mathcal{S}$
  Compute the difference loss $\mathcal{J}_\theta(X_s)$ with the parameters θ_t on X_s
  (Meta-process) Compute adapted parameters $\theta_t^{'}$ with gradient ascent: $\theta_t^{'} \leftarrow \theta_t + \eta \nabla_\theta \mathcal{J}_\theta(X_s)$
  Compute the vanilla loss $\mathcal{L}_{\theta'}(X_t)$ with the parameters $\theta_t^{'}$ on X_t
  (MetaRM-optimization) Update the parameters θ_t with gradient descent: $\theta_{t+1} \leftarrow \theta_t - \alpha \nabla_{\theta'} \mathcal{L}_{\theta'}(X_t)$
end for

meta-optimization and vanilla optimization) to iteratively adapt the RM to the new environment distribution. Our method can be summarised as the RM performs a meta-process by maximizing the difference loss function $\mathcal{J}_\theta$ before the original gradient update. Let $\mathcal{S} = \{(x^i, s^i), 1 \leq i \leq M\}$ denotes the meta dataset sampled from a shifted distribution. The meta-process can be represented as updating parameters by a gradient ascent of the difference loss function $\mathcal{J}_\theta$ on a mini-batch X_s of the meta dataset $\mathcal{S}$. Formally, at step t of the training phase, the parameters of the RM r_θ are adjusted according to the ascent direction:

$$\theta_t^{'} = \theta_t + \eta \frac{\partial \mathcal{J}_\theta(X_s)}{\partial \theta}, \tag{3}$$

where η controls the degree of learning differences between responses from the meta dataset $\mathcal{S}$. Subsequently, we compute the gradient of the vanilla loss function $\mathcal{L}_{\theta'}$ wrt. the parameters $\theta^{'}$ of the RM on a mini-batch $X_t = \{(x^i, y_w^i, y_l^i), 1 \leq i \leq n\}$ of the original preference pairs dataset $\mathcal{D}$, which can be represented as follows:

$$\nabla\theta = \frac{\partial \mathcal{L}_{\theta'}(X_t)}{\partial \theta'}. \tag{4}$$

The x^i in each batch X_s of the meta dataset $\mathcal{S}$ does not need to match the x^i in X_t.

Note that the MetaRM-optimization using the gradient $\nabla\theta$ is performed over the RM parameters θ, whereas the objective $\mathcal{L}_\theta$ is computed using the updated RM parameters $\theta^{'}$. Essentially, MetaRM seeks to learn more from these preference pairs, which can provide more information to differentiate between responses sampled from the shifted environment distribution. Formally, the MetaRM-optimization is performed via gradient descent, and the RM parameters θ are optimized as follows:

$$\theta_{t+1} = \theta_t - \alpha \nabla\theta, \tag{5}$$

where α is the learning rate for vanilla optimization. The full algorithm is detailed in Algorithm 1.

Analysis of Optimization Objective

To elucidate the aim of MetaRM, we derive the gradient $\nabla\theta$ (*i.e.,* Equation 4) of optimizing the reward model r_θ:

$$\begin{aligned}
\nabla\theta &= \frac{\partial \mathcal{L}_{\theta'}(X_t)}{\partial \theta'} \\
&= \frac{\partial \mathcal{L}_{\theta'}(X_t)}{\partial \theta} (\frac{\partial \theta^{'}}{\partial \theta})^{-1} \\
&= \frac{\partial \mathcal{L}_{\theta'}(X_t)}{\partial \theta} (1 + \eta \frac{\partial^2 \mathcal{J}_\theta(X_s)}{\partial \theta^2})^{-1}
\end{aligned} \tag{6}$$

where $(1 + \eta \frac{\partial^2 \mathcal{J}_\theta(X_s)}{\partial \theta^2})^{-1}$ is deterministic for X_t when the meta-dataset $\mathcal{S}$ is sampled, so it can be considered as a constant. We then apply Taylor expansion to $\mathcal{L}_{\theta'}(X_t)$ about point θ, which can be written as follows:

$$\begin{aligned}
&\mathcal{L}_{\theta'}(X_t) \\
&= \mathcal{L}_\theta(X_t) + \frac{\partial \mathcal{L}_\theta(X_t)}{\partial \theta}(\theta^{'} - \theta) + o(\theta^{'} - \theta)^2 \\
&= \mathcal{L}_\theta(X_t) + \eta \frac{\partial \mathcal{L}_\theta(X_t)}{\partial \theta} \frac{\partial \mathcal{J}_\theta(X_s)}{\partial \theta} + o(\theta^{'} - \theta)^2 \\
&= \mathcal{L}_\theta(X_t) + \eta \sum_{i=1}^{n} \frac{\partial \mathcal{L}_\theta(x_i)}{\partial \theta} \frac{\partial \mathcal{J}_\theta(X_s)}{\partial \theta} + o(\theta^{'} - \theta)^2
\end{aligned} \tag{7}$$

where o is infinitesimals that can be ignored.

Substituting Equation 7 into Equation 4, we obtain the gradient $\nabla\theta$:

$$\nabla\theta \propto \frac{\partial}{\partial \theta}[\mathcal{L}_\theta(X_t) + \sum_{i=1}^{n} \frac{\partial \mathcal{L}_\theta(x_i)}{\partial \theta} \frac{\partial \mathcal{J}_\theta(X_s)}{\partial \theta}]. \tag{8}$$

Equation 8 suggests that MetaRM-optimization essentially adds a sum of dot products to the vanilla loss function. The dot product computes the similarity between the gradient directions of the meta loss $\mathcal{J}_\theta$ wrt. θ and vanilla loss wrt. θ.

Specifically, when the direction of minimizing the vanilla loss on the preference pairs X_t and maximizing the difference between the rewards of the responses X_s are similar, the dot product of both is greater. In such instances, the gradient $\nabla\theta$ in the MetaRM-optimization is larger and the reward model r_θ can learn more about these preference pairs. Conversely, if the gradients are in different directions, these preference pairs may not be more helpful in alleviating the environment distribution shift, so we downweight the degree of optimization on these data.

Experiments

Experimental Setup

In this work, we use Llama-2 (Touvron et al. 2023) with seven billion parameters as the base model for all experiments. To evaluate the effectiveness of our method in iterative RLHF optimization, we conduct experiments on the general dialogue task and the summarization task. In addition, we also evaluate our approach in an out-of-distribution setting to demonstrate MetaRM's ability to differentiate subtle differences in OOD samples.

Generation Dialogue Task. Following Vicuna (Chiang et al. 2023), **SFT dataset** contains 52k multi-turn user-shared conversations from ShareGPT.com (ShareGPT 2023), including a variety of domains such as mathematics, knowledge querying, and coding. For **Human preference data**, we utilize Anthropic's HH-RLHF (Bai et al. 2022), a comprehensive collection of human preference concerning AI assistant responses (Bai et al. 2022). It contains 161k training samples and 8,500 testing samples including helpfulness and harmlessness data.

Summarization Task. For **SFT dataset**, we use the Reddit TL;DR dataset (Völske et al. 2017) as the training dataset, which contains 123,169 Reddit posts paired with human-authored summaries. **Human preference data** is similar to the SFT dataset, which includes preference pairs posts. Each post is paired with two generated summaries, one of which is labeled as preferred by annotators (Stiennon et al. 2020a).

Out-of-Distribution Task. SFT dataset is the same as the dataset used in the generation dialogue task. For **Human preference data**, we use the Oasst1 dataset (Köpf et al. 2024) as the helpfulness data of OOD task. This dataset is a human-annotated assistant-style conversation dataset including over 10k conversations (Köpf et al. 2023). On the other hand, we use PKU-SafeRLHF (Dai et al. 2024) as the harmlessness data, which is a human-labelled dataset containing both performance and safety preferences.

Dataset	Opponent vs SFT	GPT-4			Human		
		Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
Anthropic-Harmless	Round 1	44	44	12	48	32	20
	Round 2	65	31	4	63	28	9
	Round 3	**69**	**28**	**3**	**72**	**22**	**6**
	Round 4	64	31	5	68	27	5
Anthropic-Helpful	Round 1	39	52	9	44	39	17
	Round 2	62	33	5	65	27	8
	Round 3	**73**	**23**	**4**	**69**	**29**	**2**
	Round 4	67	27	6	65	23	12
Summary	Round 1	51	11	38	54	16	30
	Round 2	55	15	30	57	12	31
	Round 3	67	14	19	63	15	22
	Round 4	**78**	**5**	**17**	**77**	**7**	**16**
	Round 5	72	8	20	69	12	19

Table 1: Main results on iterative RLHF optimization. We compare the win, tie, and lose ratios of RLHF by MetaRM in the different rounds against the SFT model under both GPT-4 and human evaluations. The results show the superior performance of LLM alignment by using MetaRM. It also highlights the consistency between human and GPT-4 evaluations.

Baselines. Our Baseline approaches include Supervised Fine-Tuning (SFT), Proximal Policy Optimization (PPO) (Schulman et al. 2017) in RLHF (Ouyang et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2023). The detailed description is discussed in the supplementary material.

Implementation Details

In the SFT phase, the learning rate is set to $2e^{-5}$, and we train two epochs with a linear decay to zero. We employ a warmup period of 0.3 epochs. The fine-tuning process was conducted on a single node with eight Nvidia A100-80G GPUs and the global batch size is set to 32. **In the reward modelling phase**, the learning rate is set to $5e^{-6}$, and the global batch size is set to 16 for both the vanilla training phase and the meta-process phase. The training epoch on original preference pair datasets is only one for our proposed method and all baselines. For each optimization round of MetaRM, the learning rates α and η are both set to $5e^{-6}$. The meta dataset is constructed from the previous iteration round and for round 1, the responses are generated by the SFT model. We sample five responses for each prompt in the training dataset to compute Equation 2. **In the PPO phase**, the learning rate for the policy model and critic model is $5e^{-7}$ and $1.5e^{-6}$. For each query, we collect 16 roll-out samples using nucleus sampling. the temperature, top-p and the repetition penalty in the sampling phase are set to 0.8, 0.9 and 1.1, respectively. We set the token-level KL penalty coefficient β to 0.05 with a clip value of 0.8.

For the iterative RLHF process, we utilize the current policy model to sample multiple responses from the original prompt dataset to obtain the meta-data. For the OOD setting, the policy model is similarly employed to sample multiple responses from the OOD prompt dataset to obtain the meta-data.

Metrics & Evaluation

Win rate. To evaluate the effectiveness of our method, we assess it by comparing its **win rate** with other baselines. Specifically, we randomly select 100 prompts from the test datasets and generate the responses from our method and baselines, respectively. We then provide these pairs of prompts and responses to human evaluators, asking them to determine which response is of higher quality, more useful, and harmless. During the entire evaluation process, the human evaluators are unaware of the responses' sources. Additionally, some studies indicate that GPT-4's evaluation of the responses aligns closely with that of human evaluators (Chang et al. 2023; Zheng et al. 2023a,c). So we also utilize GPT-4 to evaluate the performance of MetaRM against other baselines. The GPT-4 prompts for evaluation can be found in the supplementary material.

Diversity. To evaluate the diversity of prompts generated by LLMs, we employ the SelfBLEU (Zhu et al. 2018) score to evaluate diversity in the form of text and sentence embeddings to evaluate diversity in the semantics of text (Zhu et al. 2018; Reimers and Gurevych 2019). The mathematical forms of the two diversity metrics can be found in the supplementary material. Specifically, we calculate the average SelfBLEU scores using n-grams for $n \in \{2, 3, 4, 5\}$ and normalize both metrics, with lower values indicating greater diversity (Zhu et al. 2018). The metrics are computed based on all the test set data, and the diversity of responses is defined as the sum of these two diversity metrics.

Main Results

Experimental results on iterative RLHF optimization. We iteratively optimize the LLM by recovering the reward model's distinguishing ability through MetaRM without collecting extra preference pairs. We recorded the improvement achieved by our approach in each optimization

Dataset	Opponent	GPT-4			Human		
		Win↑	Tie	Lose↓	Win↑	Tie	Lose↓
Anthropic-Harmless	SFT	69 (68.3)	28 (27.0)	3 (4.7)	72 (71.3)	22 (21.7)	6 (7.0)
	Vanilla PPO	54 (53.7)	31 (30.0)	15 (16.3)	58 (57.0)	24 (23.7)	18 (19.3)
	DPO	49 (49.3)	16 (15.0)	35 (35.7)	53 (51.0)	14 (16.7)	33 (32.3)
Anthropic-Helpful	SFT	73 (73.3)	23 (21.7)	4 (5.0)	69 (68.3)	29 (25.0)	2 (6.7)
	Vanilla PPO	65 (64.3)	30 (28.3)	5 (7.3)	67 (66.0)	28 (26.7)	5 (7.3)
	DPO	58 (58.3)	35 (32.3)	7 (9.3)	56 (54.7)	34 (31.7)	10 (13.7)
Summary	SFT	78 (77.0)	5 (6.3)	17 (16.7)	77 (75.7)	7 (10.3)	16 (14.0)
	Vanilla PPO	62 (61.7)	7 (9.0)	31 (29.3)	54 (55.0)	19 (16.3)	27 (28.7)
	DPO	59 (59.7)	6 (10.7)	35 (29.7)	66 (64.0)	14 (15.0)	20 (21.0)

Table 2: The results compare RLHF by MetaRM against the SFT model and other popular alignment baselines. The values in parentheses indicate the average values under different learning rates η in the sensitivity analysis experiment.

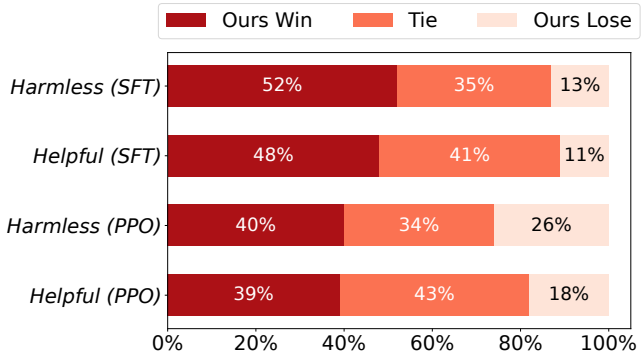

Figure 3: The results on the out-of-distribution task compared to SFT and vanilla PPO. The results show that our method outperforms other baselines by adapting the reward model to the new distribution.

round, in comparison to the SFT model, as written in Table 1. In addition, to more comprehensively demonstrate the superiority of our approach, we also compare the best round of MetaRM (*i.e.,* rounds three and four in the generation dialogue task and the summarization task, respectively) against other widely used baselines including the vanilla PPO (Ouyang et al. 2022) and DPO (Rafailov et al. 2023), as shown in Table 2.

From the results of the two tables, we can observe that: **(1)** In each round, our proposed method can significantly improve the quality of responses compared to the SFT model, both for GPT-4 and human evaluation. This improvement was notable in the initial rounds of RLHF optimization, *i.e.,* rounds one and two. **(2)** The results show a decline in the win rate in the fourth round of the dialogue generation task and the fifth round of the Summarization task. It indicates that the effectiveness of our approach has an upper limit, which varies depending on the task. **(3)** Our method significantly outperforms all other state-of-the-art baselines including the original RLHF and DPO, by iteratively training the language model without introducing extra preference pairs. **(4)** Evaluation by human evaluators aligns closely with GPT-4. Therefore, our primary reliance is placed upon the assessments from GPT-4 in subsequent experimental evaluation for saving time and resources.

Methods	Anthropic-Harmless	Anthropic-Helpful	Summary
Raw data	0.14	0.09	0.07
SFT	0.07	0.04	0.06
Vanilla PPO	0.47	0.29	0.36
DPO	**0.39**	0.31	0.38
$\text{RLHF}_{\text{MetaRM}}$	0.41	**0.25**	**0.33**

Table 3: The diversity results compare RLHF by MetaRM against raw good responses and other baselines. Lower values indicate greater diversity.

Experimental results on out-of-distribution task. We also apply MetaRM in an OOD setting to demonstrate its ability to adapt the reward model to a new out-of-distribution. The experimental results are shown in Figure 3. The results reveal that MetaRM can enhance the performance of LLM alignment in the OOD task. MetaRM can increase the RM's ability to identify subtle differences in responses sampled from OOD prompts to improve its performance in the RL training phase without extra preference data. The outstanding experimental results highlight the effectiveness and potential of our framework for LLM alignment in both ID and OOD scenarios.

Experimental results on diversity of responses. Recently, researchers found that the diversity of responses decreases during the alignment phase. To evaluate the response diversity of the LLM optimized by RLHF with MetaRM, we use two diversity metrics to access our proposed method and other alignment baselines on the test set, as shown in Table 3. Experimental results show that the diversity of responses generated by the SFT model is better than the diversity of good responses in the preference pairs. All LLM

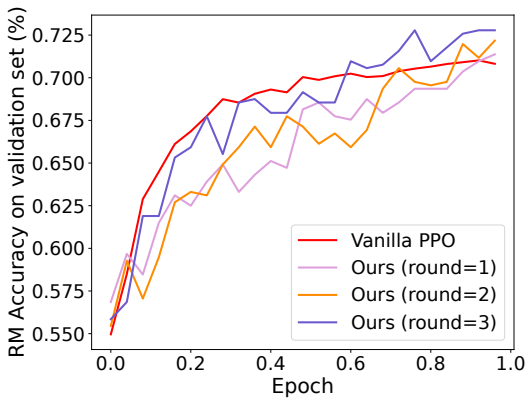

Figure 4: The accuracy curves in the reward model training phase on the valid set. The curves show that MetaRM can achieve similar accuracy compared to the original RM training way. This indicates that our method can preserve the RM's ability to modeling human preferences in the gradient descent, while making it adapt to the new distribution by using the meta-process.

alignment approaches reduce the diversity of responses, although they improve the helpfulness and the harmlessness of responses. On the other hand, compared to other alignment methods, our method can slightly increase the diversity of responses on helpfulness and summary tasks. The results indicate that our proposed method can improve the quality of responses iteratively, while slightly increasing the diversity of responses.

Sensitivity Analysis

Compared to the vanilla reward modeling method, MetaRM introduces another hyper-parameter η to control the degree of learning differences between samples of the meta dataset, as shown in Equation 3. To further evaluate the effectiveness of MetaRM, we analyze the hyper-parameter impact. Specifically, we set η to $1e^{-6}$, $5e^{-6}$, and $1e^{-5}$, respectively, and fix other hyper-parameters to train the reward model. As with the previous experimental setting, we select the LLM of round three and round four in the generation dialogue task and the summarization task, respectively, and compare them with other alignment methods by GPT-4 and Human annotation. The average comparison results are shown in parentheses of Table 2. The results reveal that MetaRM can significantly improve the performance of LLM alignment across different experimental settings. Additionally, the performance of MetaRM does have little fluctuation across different degrees of learning differences in shifted distribution, which demonstrates the stability of our proposed method.

Discussion

The Accuracy curves in the RM training phase. We record the reward model accuracy curves of the original RM training approach (*i.e.,* as defined by Equation 1) and several training rounds of the MetaRM way during the training phase, as shown in Figure 4. Compared to the original RM

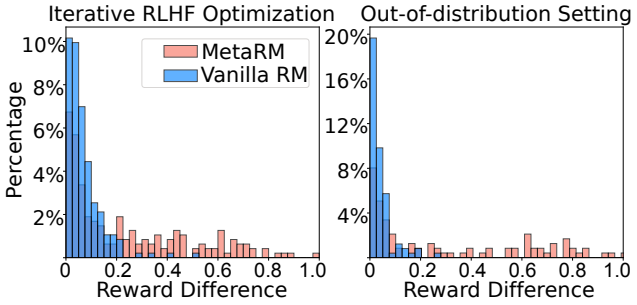

Figure 5: Reward difference distributions for the original RM's training way and MetaRM, which normalize to a range of zero to one. It indicates that MetaRM can enhance the RM's ability to distinguish samples from a shifted environment distribution through meta-learning.

training way, we can observe that the MetaRM does not affect the accuracy of the reward model on the valid set of the preference dataset, although we introduce an additional gradient ascent process on the meta dataset. This indicates that our method can enhance the reward model the capability of aligning with the new environment distribution while maintaining the ability to model human preferences through meta-learning. In addition, the trend of each round's curve shows a high consistency which represents the reasonable and effectiveness of our proposed approach.

Reward Difference Distribution. We randomly select 1,000 prompts and plot the reward difference distribution of vanilla RM after PPO training and RM after MetaRM training, respectively, as shown in Figure 5. The reward difference means the absolute difference in rewards given by the RM for different responses under the same prompt. It means whether the reward model can capture the subtle differences between the samples in the new distribution. The results show that the difference generated by the reward model trained using the original RM way is centered in the range of zero to 0.2. On the contrary, the difference given by the RM trained using MetaRM exhibits lower peaks and greater dispersion. This indicates that our method significantly enhances the RM's ability to distinguish responses sampled from a shifted environment distribution.

Conclusion

In this paper, we introduce MetaRM, a method that adapts the reward model to the shifted environment distribution through meta-learning. MetaRM iteratively trains the RM in an alternating way, to maximise discrimination power over responses of the shifted distribution, while preserving its ability to modeling human preference from the original preference pairs. Extensive experiments show that MetaRM can constantly achieve an improvement of alignment within the iterative RLHF optimization, while enhancing RM's capability of differentiating subtle differences in OOD samples.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially

funded by National Natural Science Foundation of China (No.62476061,62206057), Shanghai Rising-Star Program (23QA1400200), Natural Science Foundation of Shanghai (23ZR1403500), and Program of Shanghai Academic Research Leader under grant 22XD1401100.

References

Bahdanau, D.; Brakel, P.; Xu, K.; Goyal, A.; Lowe, R.; Pineau, J.; Courville, A.; and Bengio, Y. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; Showk, S. E.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T. B.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Bansal, T.; Jha, R.; Munkhdalai, T.; and McCallum, A. 2020. Self-supervised meta-learning for few-shot natural language classification tasks. *arXiv preprint arXiv:2009.08445*.

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

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Zhu, K.; Chen, H.; Yang, L.; Yi, X.; Wang, C.; Wang, Y.; et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Chen, Y.; Zhong, R.; Zha, S.; Karypis, G.; and He, H. 2021. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Coste, T.; Anwar, U.; Kirk, R.; and Krueger, D. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2024. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *The Twelfth International Conference on Learning Representations*.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Dou, Z.-Y.; Yu, K.; and Anastasopoulos, A. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Eschmann, J. 2021. Reward function design in reinforcement learning. *Reinforcement Learning Algorithms: Analysis and Applications*, 25–33.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, 1126–1135. PMLR.

Gao, L.; Schulman, J.; and Hilton, J. 2022. Scaling Laws for Reward Model Overoptimization. arXiv:2210.10760.

Hospedales, T.; Antoniou, A.; Micaelli, P.; and Storkey, A. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5149–5169.

Jia, C. 2024. Generalizing Reward Modeling for Out-of-Distribution Preference Learning. *arXiv preprint arXiv:2402.14760*.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z.-R.; Stevens, K.; Barhoum, A.; Duc, N. M.; Stanley, O.; Nagyfi, R.; et al. 2023. OpenAssistant Conversations–Democratizing Large Language Model Alignment. *arXiv preprint arXiv:2304.07327*.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Krueger, D.; Maharaj, T.; and Leike, J. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K.; Mesnard, T.; Bishop, C.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5051–5059.

Min, S.; Lewis, M.; Zettlemoyer, L.; and Hajishirzi, H. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774, 2023:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pace, A.; Mallinson, J.; Malmi, E.; Krause, S.; and Severyn, A. 2024. West-of-N: Synthetic Preference Generation for Improved Reward Modeling. *arXiv preprint arXiv:2401.12086*.

Pikus, B.; LeVine, W.; Chen, T.; and Hendryx, S. 2023. A Baseline Analysis of Reward Models' Ability To Accurately Analyze Foundation Models Under Distribution Shift. *arXiv preprint arXiv:2311.14743*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Ramé, A.; Vieillard, N.; Hussenot, L.; Dadashi, R.; Cideron, G.; Bachem, O.; and Ferret, J. 2024. WARM: On the Benefits of Weight Averaged Reward Models. *arXiv preprint arXiv:2401.12187*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

ShareGPT. 2023. ShareGPT. https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered. Accessed: 2024-07-01.

Shen, W.; Zheng, R.; Zhan, W.; Zhao, J.; Dou, S.; Gui, T.; Zhang, Q.; and Huang, X. 2023. Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. arXiv:2310.05199.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020a. Learning to summarize with human feedback. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 3008–3021. Curran Associates, Inc.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020b. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Völske, M.; Potthast, M.; Syed, S.; and Stein, B. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, 59–63.

Wulfe, B.; Balakrishna, A.; Ellis, L.; Mercat, J.; McAllister, R.; and Gaidon, A. 2022. Dynamics-aware comparison of learned reward functions. *arXiv preprint arXiv:2201.10081*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023a. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zheng, R.; Dou, S.; Gao, S.; Shen, W.; Wang, B.; Liu, Y.; Jin, S.; Liu, Q.; Xiong, L.; Chen, L.; et al. 2023b. Secrets of rlhf in large language models part I: Ppo. *arXiv preprint arXiv:2307.04964*.

Zheng, R.; Shen, W.; Hua, Y.; Lai, W.; Dou, S.; Zhou, Y.; Xi, Z.; Wang, X.; Huang, H.; Gui, T.; et al. 2023c. Improving Generalization of Alignment with Human Preferences through Group Invariant Learning. *arXiv preprint arXiv:2310.11971*.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 1097–1100.

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