Reinforcement Learning for Large Language Models via Group Preference Reward Shaping

Huaisheng Zhu[1], Siyuan Xu[1], Hangfan Zhang[1], Teng Xiao[1], Zhimeng Guo[1], Shijie Zhou[2], Shuyue Hu[3], Vasant G Honavar[1]
[1]Pennsylvania State University, [2]University at Buffalo,
[3]Shanghai Artificial Intelligence Laboratory
{hvz5312,spx5032,hbz5148,tengxiao,zhimeng,vuh14}@psu.edu
shijiezh@buffalo.edu, hushuyue@pjlab.org.cn

Abstract

Large Language Models (LLMs) require alignment via reinforcement learning (RL) to effectively perform task-specific objectives, such as human preference alignment and enhanced reasoning. While Proximal Policy Optimization (PPO) is widely adopted, its computational overhead, stemming from additional value model requirements, limits applicability. Existing alternatives, like Group Relative Policy Optimization (GRPO), mitigate computational costs but remain sensitive to reward model quality. To address this, we introduce Group Preference Reward Shaping (GPRS), a novel method that leverages preference-based comparisons rather than precise numerical rewards. GPRS requires no extra model components and remains robust across varying reward model sizes and qualities. Extensive experiments demonstrate that GPRS consistently outperforms existing critic-model-free RL algorithms in Reinforcement Learning from Human Feedback (RLHF) and reasoning tasks, providing stable and good alignment performance.

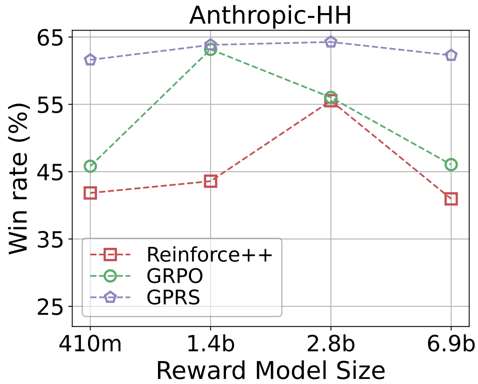

Figure 1: Performance comparison of different algorithms using the same base model (Pythia-2.8B) with regard to different size of reward models.

1 Introduction

Pre-training furnishes large language models (LLMs) with expansive world knowledge, but it does not guarantee compliance with task-specific requirements. To achieve the desired performance on particular tasks, a post-training process—often referred to as alignment or fine-tuning—is crucial. Alignment plays a pivotal role in guiding Large Language Models (LLMs) to excel at specific target attributes (e.g., helpfulness), while minimally affecting off-target attributes (e.g., reasoning ability). Notably, reinforcement learning (RL) techniques have consistently demonstrated success in aligning LLMs with human preferences and enhancing their reasoning abilities during fine-tuning (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022). These approaches reframe text generation as a sequential decision-making process, enabling precise optimization of the model's outputs to meet alignment goals, which is widely used in cutting-edge models (Achiam et al., 2023).

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is the most widely adopted reinforcement learning algorithm for fine-tuning LLMs, as reflected in both academic research and open-source implementations (Ouyang et al., 2022; Yao et al., 2023). However, PPO imposes heavy computational overhead because it requires an additional critic model and related training components to optimize this model effectively. In practice, training with PPO can take up to four times longer than both supervised fine-tuning and reward model training (the first stage in RL algorithms aligned with human preferences) and requires at least twice the GPU memory, as noted by Yao et al. (2023). This substantial overhead makes RL prohibitive for projects with limited computational resources.

To address these challenges, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has been introduced as a variant of PPO that eliminates the critic model in favor of group based normalization, thereby significantly reducing training resources. Although GRPO has shown promise on reasoning tasks, its reward normalization strategy

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21387–21400
November 4-9, 2025 ©2025 Association for Computational Linguistics

primarily focuses on variance reduction and heavily relies on reward models fine-tuned from human preference data. This is because these reward models may not precisely capture the true reward and have noise unlike reasoning tasks where correctness rewards can be explicitly measured. The training performance of GRPO can vary significantly and is highly sensitive to the size and quality of the reward model, as illustrated in Figure 1.

To address this issue, we leverage the unique characteristics of Reinforcement Learning from Human Feedback (RLHF), recognizing that humans often cannot assign an exact score to a single LLM-generated response but can readily compare responses by preference. Building on this insight, we propose a Group Preference Reward Shaping (GPRS) approach. Specifically, we reshape the rewards by comparing the current model's responses with those from the previous step, using scores provided by the reward model. Moreover, we replace comparisons across optimization steps with a simpler approach: reshaping rewards based on comparisons within response groups. As shown in Figure 1, this method outperforms group normalization in RLHF tasks and achieves consistent performance across reward models of different sizes.

Our contributions. We propose Group Preference Reward Shaping (GPRS), an efficient algorithm that requires no additional model components during training and remains robust to variations in reward models—typically trained on human preference data, which aligns well with our group preference shaping methods. Rather than relying on potentially noisy numeric scores, our approach employs preference-based comparisons. Additionally, we provide theoretical analysis showing that GPRS converges to a state where current responses consistently achieve higher win rates throughout training.

Experiments. We conduct extensive experiments with GPRS on RLHF tasks and observe consistent improvements over existing RL algorithms that do not require critic model training. Our results also show that GPRS remains robust to variations in reward models, surpassing other critic-model-free algorithms. Finally, we find that GPRS has comparable or even better results with group normalization algorithms on reasoning tasks.

2 Related Works

Alignment has become a key method for improving the performance of pre-trained language models, especially in complex instruction-following tasks like commonsense reasoning, coding, summarization, and math problem solving (Bai et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020; Rafailov et al., 2023). In this context, diverse reinforcement learning approaches have been introduced to align large language models with human preferences or to enhance their reasoning capabilities (Casper et al., 2023; Chaudhari et al., 2024; Yao et al., 2023).

In particular, RL-based alignment has attracted significant interest in the quest for safer and more helpful LLMs known as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Stiennon et al., 2020; Rafailov et al., 2023; Xiao et al., 2024b,a, 2025a,b). In this framework, the first step employs supervised fine-tuning to provide the model with a solid initialization, the second step involves training a reward model—often with Bradley-Terry (BT) models (Bradley and Terry, 1952)—and the final step leverages RL algorithms. Moreover, a lot of research has shown that RL with role-based rewards without training the reward models can achieve great results on mathematical reasoning tasks (Chaudhari et al., 2024; Hu, 2025; Yang et al., 2024; Kumar et al., 2024; Zeng et al., 2025). Notably, Proximal Policy Optimization (PPO) algorithms (Schulman et al., 2017) have been applied extensively to RLHF and reasoning tasks across a range of settings. However, PPO typically relies on training an additional critic model, causing substantial computational overhead during fine-tuning. While various techniques have been introduced to reduce the memory footprint of PPO algorithms (Sohoni et al., 2019; Rajbhandari et al., 2020; Zheng et al., 2023), they still incur the computational burden associated with PPO's critic model. To alleviate this overhead, methods such as Group Relative Policy Optimization (GRPO) (Yao et al., 2023) and Reinforce++ (Hu, 2025) have been proposed, which remove or streamline the critic model from the training process. However, these approaches are designed principally for reducing reward variance via normalization (e.g., group normalization or batch normalization) and, as illustrated in Figure 1, can be unstable when applied to reward models of varying sizes on RLHF tasks.

In practice, RLHF reward models often rely on pairwise preference data (BT models), but using their outputs as exact numeric scores to train RL algorithms can lead to inaccuracies. Reasoning tasks remain unaffected because correctness can directly serve as the reward. To address this discrepancy, we introduce a group preference reward shaping method that compares groups of responses according to the score given by reward models. This design closely mirrors human evaluation and aligns with reward model objectives, yielding robust results in RLHF tasks while also preserving strong performance on reasoning tasks.

3 Preliminaries and Notations

Let the text sequence $\mathbf{x} = [x_1, x_2, \ldots]$ represent the input prompt, and $\mathbf{y} = [y_1, y_2, \ldots]$ represent the generated response. We denote the policy by $\pi_\theta(\mathbf{y} \mid \mathbf{x})$, where θ are the model parameters and the policy defines the probability of generating response $\mathbf{y}$ conditioned on the input $\mathbf{x}$. Specifically, given a context $\mathbf{x}$, the LLM models the generation process autoregressively, where at each time step t, a token is sampled as $\mathbf{y}_t \sim \pi_\theta(y_t \mid y_1, \ldots, y_{t-1}, \mathbf{x})$. This process continues until an end-of-sentence (EOS) token is generated or a predefined maximum length T is reached. In this paper, we consider the problem of reinforcement learning (RL) for fine-tuning large language models (LLMs). Typically, reinforcement learning involves an initial supervised fine-tuning phase to initialize model parameters, followed by optimizing the RL objectives using reward.

RL for fine-tuning LLMs.Typically, a reward function $r(\mathbf{x}, \mathbf{y})$ is provided, which reflects human preferences in RLHF tasks or rule-based feedback—such as correctness or incorrectness of the generated response $\mathbf{y}$—for reasoning tasks. To fine-tune large language models (LLMs) using a reward function, a commonly adopted approach is Proximal Policy Optimization (PPO) (Schulman et al., 2017)—a widely used actor-critic Reinforcement Learning algorithm (Ouyang et al., 2022). PPO is particularly popular in the RL fine-tuning stage of LLMs, where it optimizes the model by maximizing the following surrogate objective:

$$\begin{aligned}\mathcal{L}_{\text{PPO}} = \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} & \left[A\left(\mathbf{x},\mathbf{y}\right)\min\{\right. \\ & \left. f_\theta\left(\mathbf{x},\mathbf{y}\right), \text{clip}\left(f_\theta\left(\mathbf{x},\mathbf{y}\right), 1-\epsilon, 1+\epsilon\right)\}\right],\end{aligned} \tag{1}$$

where we simplify $\mathbf{x} \sim \rho(\mathbf{x})$ and $\mathbf{y} \sim \pi_{\theta_{\text{old}}}(\mathbf{y} \mid \mathbf{x})$ as $\mathbf{x} \sim \rho$ and $\mathbf{y} \sim \pi_{\theta_{\text{old}}}$. θ_{old} is the parameter of the old policy model. Notably, training trajectories $\mathbf{y}$ are sampled from an older policy $\pi_{\theta_{\text{old}}}$, and importance sampling is employed to correct for the distribution shift. This is done using the importance weight ratio: $f_\theta(\mathbf{x}, \mathbf{y}) = \pi_\theta(\mathbf{y} \mid \mathbf{x})/\pi_{\theta_{\text{old}}}(\mathbf{y} \mid \mathbf{x})$, which helps reduce bias during optimization. ϵ is a clipping hyperparameter used to stabilize the training process. The advantage function $A(\mathbf{x}, \mathbf{y})$ quantifies how much better an action $\mathbf{y}$ is compared to the expected outcome at prompt $\mathbf{x}$, and is defined as $A(\mathbf{x}, \mathbf{y}) = r(\mathbf{x}, \mathbf{y}) + V(\mathbf{x}, \mathbf{y}) - V(\mathbf{x})$, where $r(\mathbf{x}, \mathbf{y})$ is the reward and V denotes the value model. The value model $V(\mathbf{x})$ is trained using a Temporal Difference (TD) learning objective (Sutton, 1988) to estimate the expected long-term return from prompt $\mathbf{x}$. Moreover, in PPO, a value function is trained jointly with the policy model. To prevent over-optimization of the reward model, a standard practice is to incorporate a Kullback–Leibler (KL) divergence penalty between the current policy and a fixed reference model. This KL term is added to the reward at each token to regularize the learning process:

$$r(\mathbf{x}, \mathbf{y}) = r_\phi(\mathbf{x}, \mathbf{y}) - \beta \log \frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}, \tag{2}$$

where ϕ denotes the parameters of the reward model, π_{ref} represents the reference policy (the initial SFT model) and β is the weighting coefficient for the KL divergence penalty. Even though PPO is effective in fine-tuning LLMs, it introduces a (trainable) value model V to load and optimize.

GRPO. More recently, to eliminate the need for training an additional value model in PPO, GRPO (Shao et al., 2024) has been proposed. It uses the average reward of multiple sampled responses to the same prompt $\mathbf{x}$ as a baseline. Specifically, GRPO firstly samples a set of K responses $\{\mathbf{y}^1, \ldots, \mathbf{y}^K\}$ from the old policy $\pi_{\theta_{\text{old}}}$, which is denotes as $\mathcal{Y}$. Based on the group of responses, GRPO introduces a slight modification to the original PPO objective shown as follows:

$$\begin{aligned}\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\mathbf{x}\sim\rho,\mathcal{Y}\sim\pi_{\theta_{\text{old}}}} & \Big[\sum_{\mathbf{y}\in\mathcal{Y}} \frac{A'\left(\mathbf{x},\mathbf{y}\right)}{K}\min\{ \\ & f_\theta\left(\mathbf{x},\mathbf{y}\right), \text{clip}\left(f_\theta\left(\mathbf{x},\mathbf{y}\right), 1-\epsilon, 1+\epsilon\right)\} \\ & -\beta\mathbb{D}_{KL}\left[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\theta_{\text{ref}}}(\mathbf{y} \mid \mathbf{x})\right]\Big],\end{aligned} \tag{3}$$

where A' denotes the advantage, which is computed based solely on the relative rewards within each group of sampled outputs from the same prompt. For each prompt $\mathbf{x}$, the corresponding rewards form a set $\mathbf{R} = \left[r(\mathbf{x}, \mathbf{y}^1), \ldots, r(\mathbf{x}, \mathbf{y}^K)\right]$. To get $\hat{A}$, GRPO normalizes these rewards using their mean and standard deviation, thereby stabilizing training and ensuring consistency across samples:

$$A'\left(\mathbf{x}, \mathbf{y}\right) = \frac{r(\mathbf{x}, \mathbf{y}) - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}, \tag{4}$$

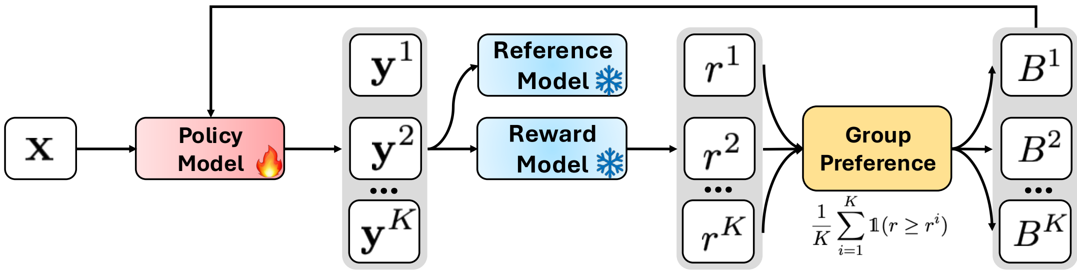

Figure 2: The illustration of model framework for Group Preference Reward Shaping (GPRS) method.

where mean and std are the mean value and standard deviation of rewards within groups. GRPO has demonstrated strong effectiveness, initially on mathematical reasoning tasks, and has since proven to be effective across a broader range of tasks (Chen et al., 2025; Liu et al., 2025; Li et al., 2025).

4 The Proposed Method

We introduce Group Preference Reward Shaping (GPRS), a lightweight yet powerful refinement of PPO. GPRS replaces noisy absolute reward signals with pairwise preference rewards obtained by comparing responses generated for the same prompt, which mirrors how humans judge outputs in RLHF. This modification makes the algorithm far more robust to imperfect reward models trained on preference data. We detail a simple implementation that only minimally alters standard PPO without critic models, and we provide theoretical analysis about GPRS , which guarantees monotonic improvements in win-rate, the key metric for evaluating RLHF systems. The training framework of GPRS is shown in Figure 2.

4.1 Group Preference Reward

GRPO normalizes rewards across a group of responses, effectively reducing the variance in the training process and serving as a substitute for the critic model used in PPO-based reinforcement learning algorithms. However, while this normalization improves stability, it is primarily designed for variance reduction and remains sensitive to the quality of the reward model, as illustrated in Figure 1. Before analyzing the failure cases associated with different reward models, we begin by examining how reward models are typically trained for human preference alignment. A common approach is to use the Bradley-Terry (BT) model (Bradley and Terry, 1952), which is trained on human preference data by optimizing the following objective:

$$p\left(\mathbf{y}_1 \succ \mathbf{y}_2 \mid \mathbf{x}\right) = \sigma\left(r\left(\mathbf{x}, \mathbf{y}_1\right) - r\left(\mathbf{x}, \mathbf{y}_2\right)\right), \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid function, $\mathbf{y}_1$ is the human-preferred response, and $\mathbf{y}_2$ is the human-dispreferred response. The trained reward model $r(\cdot)$ is then used in reinforcement learning algorithms such as those in Equations (1), (3). Based on this formulation, we attribute the failures observed with different reward models to the fact that the trained reward may not accurately reflect the true reward values. Instead, reward model training primarily captures relative human preferences rather than absolute reward magnitudes, which can result in inaccurate absolute reward values. Consequently, the noise introduced by these inaccurate rewards may lead to suboptimal performance. Additionally, human evaluation practices commonly adopted for RLHF in LLMs typically rely on expressing preferences between responses rather than assigning numerical values, which highlights a misalignment between absolute rewards and the evaluation process in RLHF tasks. Motivated by this insight, we propose to optimize a preference-aligned reward, with the objective of encouraging the current policy to generate responses that are preferred over those produced by the previous policy:

$$r_p(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\mathbf{y}' \sim \pi_{\theta_{\text{old}}}}\left[r(\mathbf{x}, \mathbf{y}) \geq r\left(\mathbf{x}, \mathbf{y}'\right)\right], \quad (6)$$

where given a response $\mathbf{y}$ generated by the current policy π_θ, the objective evaluates whether this response is preferred over responses generated by the previous policy $\pi_{\theta_{\text{old}}}$. In essence, it measures the relative improvement of the current policy compared to earlier optimization steps. We utilize the reward model $r(\cdot)$, trained from human preference data as described in Equation (5), to compare two responses. Instead of providing exact reward values, the reward formulation in Equation (6) utilizes relative preferences, which better aligns with its training objective and human evaluation process. As a result, this preference-based reward offers a more accurate and reliable signal compared to the absolute rewards used in previous reinforcement learning algorithms, e.g., GRPO, which also be empirically verified in Figure 1.

4.2 Training of GPRS

In the previous section, we introduced a preference-based reward formulation. However, computing the reward $r(\cdot)$ requires sampling from the previous policy $\pi_{\theta_{\text{old}}}$, which introduces additional storage overhead for both the samples and the parameters of θ_{old}. To address this limitation, we propose an alternative approach: instead of comparing responses from the current and previous policies, we compare multiple responses sampled solely from the current policy. Specifically, given a prompt $\mathbf{x}$, we obtain a set of responses $\mathcal{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^K\}$, then we calculate the advantage with this group preference reward for each response $\mathbf{y} \in \mathcal{Y}$:

$$B(\mathbf{x}, \mathbf{y}) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{1}(r(\mathbf{x}, \mathbf{y}) \geq r(\mathbf{x}, \mathbf{y}^i)), \quad (7)$$

where we compare preference-based rewards within groups of responses. In this paper, for RLHF tasks, we use outcome rewards trained on human preference data. For reasoning tasks, the reward corresponds to the correctness of the final answer, which is assigned as 1 if correct and -1 otherwise. With this new group preference-based reward shaping strategy, we maximize the following objective using group-based preference models:

$$\begin{aligned} \mathcal{L}_{\text{GPRS}} = \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta_{\text{old}}}} [\sum_{\mathbf{y}\in\mathcal{Y}} \frac{B(\mathbf{x}, \mathbf{y})}{K} \min\{ \\ f_\theta(\mathbf{x}, \mathbf{y}), \text{clip}(f_\theta(\mathbf{x}, \mathbf{y}), 1-\epsilon, 1+\epsilon)\} \\ -\beta \mathbb{D}_{KL}[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})]]. \end{aligned} \quad (8)$$

Instead of comparing responses sampled from previous policy, GRPS only requires sampling a group of responses from the current policy. Therefore, our proposed GRPS algorithm incurs no extra computational cost compared to GRPO. To further reduce training variance, following Ouyang et al. (Ouyang et al., 2022), we center the baseline within each group as $B(\mathbf{x}, \mathbf{y}) = B(\mathbf{x}, \mathbf{y}) - \text{mean}(\mathcal{B})$, where $\mathcal{B} = \{B(\mathbf{x}, \mathbf{y}^1), \ldots, B(\mathbf{x}, \mathbf{y}^K)\}$.

4.3 Theoretical Analysis

In this section, we first introduce the theoretical insight of the algorithm designed in Section 4.2, and then provide theoretical guarantees.

Following PPO (Schulman et al., 2017), maximizing the objective in (8) is to approximate the solution of the following optimization problem:

$$\begin{aligned} \max_{\pi_\theta} \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta_{\text{old}}}} [B(\mathbf{x}, \mathbf{y}) f_\theta(\mathbf{x}, \mathbf{y}) \\ - \alpha \mathbb{D}_{KL}[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y} \mid \mathbf{x})] \\ -\beta \mathbb{D}_{KL}[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})]]. \end{aligned} \quad (9)$$

Specifically, as indicated in PPO, (8) employs a clipping function to penalize the divergence between π_θ and $\pi_{\theta_{\text{old}}}$ in (9), eliminating the need to manually select the hyperparameter α and thereby maintaining the simplicity of the training process. We next present a theoretical analysis of the optimization problem (9), providing theoretical motivation for the algorithm developed in Section 4.2.

We start by studying the property of the group preference reward function $B(\mathbf{x}, \mathbf{y})$. Let $Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}$ denote the cumulative distribution function of $r(\mathbf{x}, \mathbf{y}')$, where $\mathbf{y}' \sim \pi_{\theta_{\text{old}}}$. We have the following proposition based on the Glivenko-Cantelli theorem (Van der Vaart, 2000).

Proposition 1 *Denote that B_K is the function B in (7) when the sampling number is K, then*

$$\sup_{\mathbf{y}} \left| B_K(\mathbf{x}, \mathbf{y}) - Q_{\mathbf{x}}^{\pi_{\theta_{old}}}(\mathbf{y}) \right| \xrightarrow{a.s.} 0.$$

Proposition 1 shows that, when the sampling number K goes to infinity, the group preference reward function B_K converges to the cumulative distribution function of the reward on $\pi_{\theta_{\text{old}}}$ almost surely. Therefore, the problem (9) approximately takes $Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})$ as the reward function. Next, we show the performance guarantee of problem (9).

Theorem 1 *Consider the solution of problem (9) is denoted as π_{θ^*}. When the sampling number K goes to infinity and $\beta = 0$, the following inequality holds: $\mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta^*}}[Q_{\mathbf{x}}^{\pi_{\theta_{old}}}(\mathbf{y})] \geq \frac{1}{2}$.*

The proof of Theorem 1 is shown in Appendix A.1. Theorem 1 shows the expected reward of the one-step GPRS in (9), i.e., $\mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta^*}}[Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})]$, increases under the KL divergence term $\mathbb{D}_{KL}[\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y} \mid \mathbf{x})]$. Specifically, as shown in the proof of Theorem 1, we have $\mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta^*}}[Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \geq \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta_{\theta_{\text{old}}}}}[Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] = \frac{1}{2}$.

Note that the objective function in (9) includes three terms: (i) the first term $B(\mathbf{x}, \mathbf{y}) f_\theta(\mathbf{x}, \mathbf{y})$, which is designed to maximize the expected GPRS reward; (ii) the KL divergence between the initial policy $\pi_{\theta_{\text{old}}}$ and the optimized policy π_θ, which prevents overly large optimization step; (iii) the KL divergence between the reference model π_{ref} and the optimized policy π_θ, which is designed to limit deviation from the reference model and ensure the optimized policy remains aligned with it. In practice, the hyperparameter β is small. Therefore, Theorem 1 shows that the expected reward can be improved even if the KL regularization term constrains the optimization step.

Corollary 1 *Under the condition of Theorem 1, the inequality holds:*

$$\mathcal{P}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta^*},\mathbf{y}'\sim\pi_{\theta_{old}}}[r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}')] \geq \frac{1}{2}.$$

Corollary 1 shows the intuition of the increased expected reward in GPRS. In particular, the win rate of the optimized policy π_{θ^*} with respect to the initial policy $\pi_{\theta_{\text{old}}}$ is larger than $\frac{1}{2}$, i.e., the policy is improved after each RL training step.

Theorem 1 and Corollary 1 guarantees the policy improvement in the win rate of π^{θ^*} with respect to $\pi_{\theta_{\text{old}}}$. Next, we derive the guarantee that the win rate with respect to the reference model π_{ref} during the GPRS steps is monotonically improved.

Theorem 2 *Under the condition of Theorem 1, when $B_K(\mathbf{x},\mathbf{y})$ is replaced by*

$$\frac{1}{K}\sum_{i=1}^{K}\frac{\pi_{ref}(\mathbf{x},\mathbf{y}^i)}{\pi_{\theta_{old}}(\mathbf{x},\mathbf{y}^i)}\mathbb{1}(r(\mathbf{x},\mathbf{y}) \geq r(\mathbf{x},\mathbf{y}^i)) \quad (10)$$

the following inequalities hold:

$$\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta^*}}[Q_{\mathbf{x}}^{\pi_{ref}}(\mathbf{y})] \geq \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{old}}}[Q_{\mathbf{x}}^{\pi_{ref}}(\mathbf{y})]$$

where $Q_{\mathbf{x}}^{\pi_{ref}}$ is the cumulative distribution function of $r(\mathbf{x},\mathbf{y}')$ under $\mathbf{y}' \sim \pi_{ref}$, and

$$\begin{aligned}&\mathcal{P}_{\mathbf{x}\sim\rho,\ \mathbf{y}'\sim\pi_{ref}}[r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}') \mid \mathbf{y}\sim\pi_{\theta^*}]\\ &\geq \mathcal{P}_{\mathbf{x}\sim\rho,\ \mathbf{y}'\sim\pi_{ref}}[r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}') \mid \mathbf{y}\sim\pi_{\theta_{old}}].\end{aligned}$$

The proof of Theorem 2 is shown in Appendix A.3. Note that in Theorem 2, the expected GPRS reward in problem (9) is replaced by (10), which is to approximate $Q_{\mathbf{x}}^{\pi_{\text{ref}}}(\mathbf{y})$ by importance sampling (Robert et al., 1999). Theorem 2 indicates that, the win rate of the optimized policy π_{θ^*} with respect to the reference model π_{ref} is better than that of the initial policy $\pi_{\theta_{\text{old}}}$. Denote the policy solved at the t-step of GPRS as π_{θ_t}, the win rate of π_{θ_t} with respect to π_{ref} is larger than that of $\pi_{\theta_{t-1}}$ for any t, i.e., $\mathcal{P}[r(\pi_{\theta_t}) > r(\pi_{\text{ref}})] \geq \mathcal{P}[r(\pi_{\theta_{t-1}}) > r(\pi_{\text{ref}})] \geq \mathcal{P}[r(\pi_{\theta_{t-2}}) > r(\pi_{\text{ref}}))] \geq \cdots \geq \mathcal{P}[r(\pi_{\text{ref}}) > r(\pi_{\text{ref}})] = \frac{1}{2}$, which achieves monotonic improvement of the win rate.

5 Experiment

In this section, we present the main experimental results, highlighting the superior performance of GPRS on RLHF and reasoning tasks and its robustness to different size of reward models.

5.1 Experimental Setup

Datasets. We evaluate our methods on widely used datasets for both RLHF and reasoning tasks. Specifically, for RLHF, we use the Reddit TL;DR summarization dataset (Völske et al., 2017) and the Anthropic-HH dataset (Bai et al., 2022). For reasoning tasks, we follow the experimental setup from (Zeng et al.), fine-tuning models on training samples of MATH (Hendrycks et al., 2024). Details of the datasets are provided in Appendix B.1.

Models. For RLHF tasks, we conduct experiments using Llama3-8B (Grattafiori et al., 2024) and Pythia-2.8B (Biderman et al., 2023) as our primary models. We further analyze model sensitivity to reward model size by evaluating performance with Pythia models of varying scales: 410M, 1.4B, 2.8B, and 6.9B. For reasoning tasks, we fine-tune Qwen2.5-Math-7B (Yang et al., 2024).

Baselines. We compare our model with baseline methods that also eliminate the need for critic models, such as GRPO (Shao et al., 2024) and Reinforce++ (Hu, 2025), on both RLHF and reasoning tasks. Additionally, for reasoning tasks, we compare GPRS with models fine-tuned through SFT as well as with larger-scale LLMs.

Evaluation and Implementation Details. We evaluate performance on reasoning tasks by the accuracy of final answer for each mathmatical problems. For RLHF tasks, following the experimental setup from Yao et al. (2023), we allocate 20% of the data for supervised fine-tuning (SFT) and 40% for training the reward model. The remaining data is used for reinforcement learning. To assess performance, we employ GPT-4o-mini to compare responses from baseline or trained models against those from the SFT model, using win rate as the evaluation metric. Details are in Appendix B.2.

5.2 Comparison on RLHF tasks

Table 1 compares the performance of GPRS with other RL methods that do not rely on critic models, evaluated on the TL;DR Summarization and Anthropic-HH datasets. We conduct experiments using both Llama3-8B and Pythia-2.8B as backbone models. We firstly observe that GPRS con-

Table 1: Win rates computed by GPT-4o-mini against the SFT generated texts and on the TL;DR summarization and Anthropic-HH datasets. Best results are highlighted in **boldface**.

Dataset (→)	TL;DR Summarization			Anthropic-HH		
Method (↓) / Metric (→)	Llama3-8B	Pythia-2.8B	Average	Llama3-8B	Pythia-2.8B	Average
REINFORCE++	51.06	65.56	58.31	65.35	55.54	60.44
GRPO	52.50	57.58	55.04	55.14	56.03	55.59
GPRS	**59.57**	**66.47**	**63.02**	**65.73**	**64.27**	**65.00**

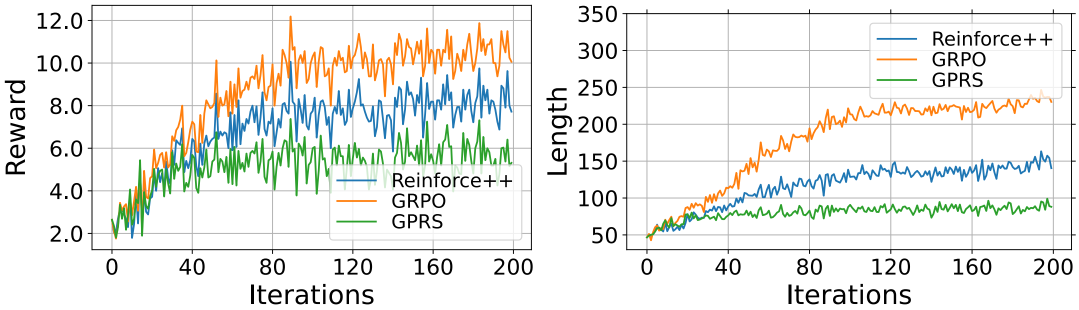

Figure 3: Training Dynamics of reward values and length of responses on Llama3-8B for Anthropic-HH.

sistently outperforms the base model, achieving win rates above 50% across all settings. This verifies our motivation to design a reward shaping approach based on preference-based reward. Moreover, our models consistently outperform both Reinforce++ and GRPO across the two evaluated datasets, further demonstrating the effectiveness of our approach using group preference-based rewards. A potential reason is that the reward models are trained on binary comparison tasks (e.g., using BT models' objectives), which are more aligned with relative preference comparisons rather than providing accurate scalar reward values. Thus, RL methods that directly optimize scalar rewards without comparing responses may be misled by the inaccuracies in these values. In contrast, our method uses relative comparison during training, aligning more closely with the reward model's objective.

We further investigate the training dynamics of raw (pre-transformation) reward values and the response lengths across different algorithms. The results are presented in Figure 3. We observe that methods such as Reinforce++ and GRPO, which employ normalization techniques for reward shaping, closely fit the reward model and significantly increase the absolute reward values. This supports our earlier observation that these methods are highly sensitive to the reward model. However, despite their ability to boost reward scores, they tend to suffer from reward hacking primarily by generating excessively long responses. This phenomenon has also been noted in prior work (Singhal et al., 2024). In our empirical analysis, we further find that while these methods lead to increased rewards and longer outputs, they often do so by generating repetitive or uninformative content. This behavior ultimately harms model performance, especially with continued training over more epochs. In contrast, our proposed GPRS not only increases the absolute reward values but also improves preference-based comparisons within groups of responses. Its training process remains stable and does not suffer from overfitting to spurious features—such as response length—often exploited by reward models.

5.3 Comparison on Reasoning Tasks

We compare the performance of models trained with GPRS against other RL algorithms, using the same base models for fine-tuning, and additionally compare GPRS against larger models with more parameters. First, we observe that RL algorithms significantly outperform SFT-based approaches, as shown in the third line. Moreover, RL fine-tuning notably enhances or even activates the reasoning abilities of smaller models (such as 7B models) with our RL-trained models surpassing the performance of Qwen-2.5-14B and LLaMA-3.1-70B-Instruct. These results validate both the effectiveness of RL algorithms and the motivation behind our work to further advance research in this area. Interestingly, our group preference reward shaping approach is designed based on preference signals, aligning closely with the human evaluation process, and achieves comparable results to normalization-based methods such as Reinforce++ and GRPO. This demonstrates that preference-based rewards are not only effective for modeling RLHF tasks but can also activate reasoning abilities, while achiev-

Table 2: Accuracy results for reasoning tasks. Best results are highlighted in **boldface**.

Models	AIME24	MATH500	AMC	Minerva Math	OlympiadBench	Average
Qwen2.5-Math-7B-Base	16.7	52.4	52.5	12.9	16.4	30.2
Qwen-2.5-Math-7B-Instruct	13.3	**79.8**	50.6	34.6	40.7	43.8
Qwen2.5-Math-7B-Base-SFT	3.3	54.6	22.5	32.7	19.6	26.5
rStar-Math-7B	**26.7**	78.4	47.5	-	47.1	-
Qwen-2.5-14B	6.7	65.4	37.5	24.3	33.5	33.5
Llama-3.1-70B-Instruct	16.7	64.6	30.1	35.3	31.9	35.7
REINFORCE++	16.7	73.6	60.0	33.5	36.0	44.0
GRPO	20.0	76.2	62.5	**38.6**	37.5	47.0
GPRS	16.7	76.4	**67.5**	**38.6**	**37.8**	**47.4**

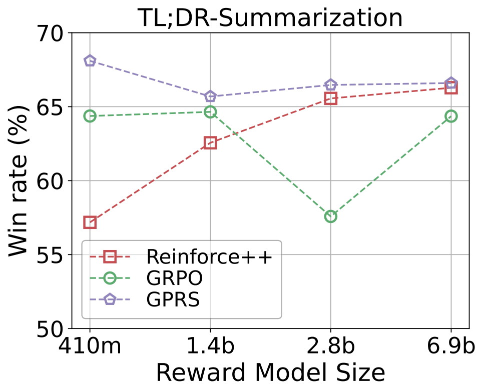

Figure 4: Performance comparison of different algorithms using the same base model (Pythia-2.8B) with regard to different size of reward models.

ing similar performance to normalization-based approaches on reasoning tasks with simple +1 (correct answer) and -1 (wrong answer) rewards. Therefore, our proposed group preference reward shaping provides a promising alternative to traditional normalization techniques for reward shaping.

5.4 Sensitivity to Reward Models

In this section, we investigate how different RL algorithms perform when trained with reward models of varying sizes. Specifically, we use the same base model, Pythia-2.8B, and fine-tune it with RL algorithms using reward models of different scales: Pythia-410M, Pythia-1.4B, and Pythia-6.9B. The corresponding results are presented in Figure 1 and Figure 4. We observe that methods relying on absolute values of rewards, such as GRPO and Reinforce++, exhibit significant sensitivity to the size of the reward model, leading to unstable performance. In contrast, GPRS demonstrates robustness across different reward model sizes. These findings support our intuition that preference-based rewards more accurately reflect the human evaluation process and the true reward training signal, resulting in more stable and reliable training outcomes.

5.5 Ablation Study

In this section, we train the model using Equation (10), which provides a theoretical guarantee of consistently improving the win rate performance over the initial SFT model. We refer to this weighted version as GPRS-weight. The results, presented in Table 3, show that GPRS achieves comparable performance to GPRS-weight. This may be attributed to the fact that the ratio $\pi_{\text{ref}}(\mathbf{x}, \mathbf{y}^i)/\pi_{\text{old}}(\mathbf{x}, \mathbf{y}^i)$ remains close to 1 or constant, as KL regularization prevents the optimized policy from deviating significantly from the reference policy π_{ref}. Based on this, we can safely omit the importance weight in Equation 10 while still achieving comparable results, relying solely on the advantage term, which retains the theoretical performance guarantee.

Table 3: Win rate for ablation study on Anthropic-HH.

Method (↓) / Metric (→)	Llama3-8B	Pythia-2.8B	Average
GPRS-weight	64.70	**66.13**	**65.41**
GPRS	**65.73**	64.27	65.00

6 Conclusion

In this paper, we propose GPRS, a simple and effective reward shaping method based on human preferences that eliminates the need for a critic model in PPO-based algorithms. GPRS reshapes reward signals by comparing responses generated by the current policy with those from the previous policy, aiming to consistently improve win-rate performance over successive iterations. We provide a theoretical analysis showing that models trained with GPRS achieve monotonic improvements in win rate with respect to the underlying reward model. GPRS shows strong results on RLHF tasks and exhibits robustness to various size of reward models. Notably, GPRS can also be applied to reasoning tasks, achieving comparable results to prior RL methods that don't rely on critic models.

7 Limitations

While GPRS demonstrates good empirical performance and is grounded in theoretical foundations, it still requires loading a reference model during training. Exploring approaches that eliminate the need for a reference model could further improve training efficiency. Additionally, GPRS is currently implemented as an on-policy RL algorithm that samples data from the current policy. Future work could investigate off-policy learning or leveraging offline datasets to train the model more effectively. We hope this work opens up new directions for reward shaping beyond traditional normalization techniques, offering alternative perspectives for preference-based reinforcement learning.

Acknowledge

This work was supported in part by grants from the National Science Foundation (2226025) and the National Center for Advancing Translational Sciences, and the National Institutes of Health (UL1 TR002014) and by the Center for Artificial Intelligence Foundations and Scientific Applications and the Institute for Computational and Data Sciences at Pennsylvania State University.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Zuyao Chen, Jinlin Wu, Zhen Lei, Marc Pollefeys, and Chang Wen Chen. 2025. Compile scene graphs with reinforcement learning. *arXiv preprint arXiv:2504.13617*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2024. Measuring mathematical problem solving with the math dataset, 2021. *URL https://arxiv. org/abs/2103.03874*.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*.

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

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, and 1 others. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Torl: Scaling tool-integrated rl. *arXiv preprint arXiv:2503.23383*.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint arXiv:2504.13055*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Christian P Robert, George Casella, and George Casella. 1999. *Monte Carlo statistical methods*, volume 2. Springer.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in rlhf. In *First Conference on Language Modeling*.

Nimit S Sohoni, Christopher R Aberger, Megan Leszczynski, Jian Zhang, and Christopher Ré. 2019. Low-memory neural network training: A technical report. *arXiv preprint arXiv:1904.10631*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Richard S Sutton. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44.

Aad W Van der Vaart. 2000. *Asymptotic statistics*, volume 3. Cambridge university press.

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

Teng Xiao, Mingxiao Li, Yige Yuan, Huaisheng Zhu, Chao Cui, and Vasant G Honavar. 2024a. How to leverage demonstration data in alignment for large language model? a self-imitation learning perspective. *arXiv preprint arXiv:2410.10093*.

Teng Xiao, Yige Yuan, Zhengyu Chen, Mingxiao Li, Shangsong Liang, Zhaochun Ren, and Vasant G Honavar. 2025a. Simper: A minimalist approach to preference alignment without hyperparameters. *arXiv preprint arXiv:2502.00883*.

Teng Xiao, Yige Yuan, Mingxiao Li, Zhengyu Chen, and Vasant G Honavar. 2025b. On a connection between imitation learning and rlhf. *arXiv preprint arXiv:2503.05079*.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024b. Cal-dpo: Calibrated direct preference optimization for language model alignment. *Advances in Neural Information Processing Systems*, 37:114289–114320.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, and 1 others. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, and 1 others. 2023. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*.

A Theoretical Analysis

A.1 Proof of Theorem 1

We provide the proof of Theorem 1 as follows.

Based on Proposition 1, when the sampling number K goes to infinity and $\beta = 0$, the optimization problem (9) becomes to

$$\max_{\pi_\theta} \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} \left[Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y}) \frac{\pi_\theta(\mathbf{x},\mathbf{y})}{\pi_{\theta_{\text{old}}}(\mathbf{x},\mathbf{y})} \right] - \alpha \mathbb{D}_{KL}\left[\pi_\theta(\mathbf{y}\mid\mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x})\right].$$

Denote the objective function in the above optimization problem as $g(\theta)$. We have that $g(\theta^*) \geq g(\theta_{\text{old}})$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y}) \frac{\pi_\theta(\mathbf{x},\mathbf{y})}{\pi_{\theta_{\text{old}}}(\mathbf{x},\mathbf{y})} \\
&- \alpha \mathbb{D}_{KL}\left[\pi_\theta(\mathbf{y}\mid\mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x})\right] \\
&\geq \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \\
&- \alpha \mathbb{D}_{KL}\left[\pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x})\right]
\end{aligned}$$

For the left-hand side, since the KL divergence is always larger than 0, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y}) \frac{\pi_\theta(\mathbf{x},\mathbf{y})}{\pi_{\theta_{\text{old}}}(\mathbf{x},\mathbf{y})}] \\
&\quad - \alpha \mathbb{D}_{KL}\left[\pi_\theta(\mathbf{y}\mid\mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x})\right] \\
=&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \\
&\quad - \alpha \mathbb{D}_{KL}\left[\pi_\theta(\mathbf{y}\mid\mathbf{x}) \| \pi_{\theta_{\text{old}}}(\mathbf{y}\mid\mathbf{x})\right] \\
\leq&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})].
\end{aligned}$$

The right-hand side is $\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})]$. Therefore we have

$$\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \geq \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})].$$

Finally, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y}) P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r) dr \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r} P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r) P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r') dr' dr \\
=&\frac{1}{2}.
\end{aligned}$$

Here $P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r)$ is the probability density function of $r(\mathbf{x},\mathbf{y}')$, where $\mathbf{y}' \sim \pi_{\theta_{\text{old}}}$. Therefore, we have $\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta^*}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})] \geq \frac{1}{2}$.

A.2 Proof of Corollary 1

We have

$$\begin{aligned}
&\mathcal{P}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta^*},\mathbf{y}'\sim\pi_{\theta_{\text{old}}}} [r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}')] \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r} P_{\mathbf{x}}^{\pi_{\theta^*}}(r) P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r') dr' dr \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} Q_{\mathbf{x}}^{\pi_{\pi_{\theta^*}}}(\mathbf{y}) P_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(r) dr \\
=&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta^*}} [Q_{\mathbf{x}}^{\pi_{\theta_{\text{old}}}}(\mathbf{y})]
\end{aligned}$$

Based on Theorem 1, Corollary 1 holds.

A.3 Proof of Theorem 2

Note that when $B_K(\mathbf{x},\mathbf{y})$ is replaced by

$$\frac{1}{K} \sum_{i=1}^{K} \frac{\pi_{\text{ref}}(\mathbf{x},\mathbf{y}^i)}{\pi_{\theta_{\text{old}}}(\mathbf{x},\mathbf{y}^i)} \mathbb{1}(r(\mathbf{x},\mathbf{y}) \geq r(\mathbf{x},\mathbf{y}^i)),$$

then

$$\sup_{\mathbf{y}} \left| B_K(\mathbf{x},\mathbf{y}) - Q_{\mathbf{x}}^{\pi_{\text{ref}}}(\mathbf{y}) \right| \xrightarrow{\text{a.s.}} 0.$$

Similar to the proof of Theorem 1, we have

$$\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta} [Q_{\mathbf{x}}^{\pi_{\text{ref}}}(\mathbf{y})] \geq \mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_{\theta_{\text{old}}}} [Q_{\mathbf{x}}^{\pi_{\text{ref}}}(\mathbf{y})].$$

Similar to the proof of Corollary 1, we have

$$\begin{aligned}
&\mathcal{P}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta,\mathbf{y}'\sim\pi_{\theta_{\text{old}}}} [r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}')] \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} \int_{r_{\min}}^{r} P_{\mathbf{x}}^{\pi_\theta}(r) P_{\mathbf{x}}^{\pi_{ref}}(r') dr' dr \\
=&\mathbb{E}_{\mathbf{x}\sim\rho} \int_{r_{\min}}^{r_{\max}} Q_{\mathbf{x}}^{\pi_{\pi_\theta}}(\mathbf{y}) P_{\mathbf{x}}^{\pi_{ref}}(r) dr \\
=&\mathbb{E}_{\mathbf{x}\sim\rho,\mathbf{y}\sim\pi_\theta} [Q_{\mathbf{x}}^{\pi_{ref}}(\mathbf{y})]
\end{aligned}$$

for any π_θ. Therefore, we have

$$\begin{aligned}
&\mathcal{P}_{\mathbf{x}\sim\rho,\ \mathbf{y}'\sim\pi_{\text{ref}}} [r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}') \mid \mathbf{y}\sim\pi_{\theta^*}] \\
&\geq \mathcal{P}_{\mathbf{x}\sim\rho,\ \mathbf{y}'\sim\pi_{\text{ref}}} [r(\mathbf{x},\mathbf{y}) > r(\mathbf{x},\mathbf{y}') \mid \mathbf{y}\sim\pi_{\theta_{\text{old}}}].
\end{aligned}$$

B Experimental Details

B.1 Details of Datasets

In this section, we provide detailed descriptions of datasets used in our experiments:

Anthropic-HH (Bai et al., 2022): The Anthropic Helpful and Harmless Dialogue dataset consists of 170,000 dialogues between humans and an automated assistant. Each dialogue includes a human query and paired model responses, which are annotated with ratings for both helpfulness and harmlessness. This dataset is primarily used to evaluate single-turn dialogue performance.

Reddit TL;DR summarization (Völske et al., 2017): This dataset comprises a curated collection of Reddit forum posts, specifically prepared for summarization tasks.

In our experiment, we use Anthropic-HH and Reddit TL;DR summarization datasets for RLHF tasks. And We prompt GPT-4o-mini for zero-shot pairwise evaluation (see Table 4 and 5).

Math (Hendrycks et al., 2024): We use the training samples from Zeng et al., consisting of mathematical queries and their corresponding final answers, for rule-based reward modeling in reinforcement learning. This dataset is employed for reasoning tasks in our experiments.

AIME24 [1]: This dataset features problems from the 2024 American Invitational Mathematics Examination (AIME), a renowned high school mathematics competition recognized for its exceptionally challenging problems.

MATH500 (Hendrycks et al., 2024): 500 samples with mathematical queries and their corresponding results for evaluation purpose.

AMC [2]: This dataset serves as an internal validation set during our participation in the AIMO progress prize competition.

Minerva Math (Lewkowycz et al., 2022): This dataset consists of over 200 undergraduate-level science and mathematics questions from MIT's OpenCourseWare (OCW), providing a benchmark for evaluating our model's quantitative reasoning abilities in a chain-of-thought setting beyond purely mathematical contexts.

OlympiadBench (He et al., 2024): This is an Olympiad-level bilingual multimodal scientific benchmark, comprising 8,476 problems from mathematics and physics competitions, including the Chinese college entrance exam. Each problem is accompanied by expert-level annotations that provide detailed step-by-step reasoning.

In our reasoning experiments, we use AIME24, MATH500, AMC and Minerva Math to evaluate the performance of models.

B.2 Implementation Details

For the general hyperparameter, we closely followed the configurations used in Yao et al. (2023) for supervised fine-tuning and reward model learning for RLHF tasks. Additionally, during the RL stage, we use a batch size of 128 for Pythia models and 64 for Llama models. The KL divergence weight, β, is set to 0.05 across all RL algorithms. We set the number of samples used for each response as 4 for all algorithms. Moreover, we follow the hyperparameter settings in Zeng et al. (2025) to train different RL algorithms for reasoning tasks. Specifically, we use a batch size of 128 and set KL divergence weight β as 0.01. We set the number of samples used for each response as 8 for all algorithms. We employ the Adam optimizer (Kingma, 2014) to fine-tune the parameters of the language models. The experiments are all run on 4 Nvidia A100 GPUs with BF16 precision. Moreover, we leverage the Copilot tool to assist with the implementation.

C Additional Experiments

We provide some case studies to analyze the quality of responses by comparing responses from GPRS, GRPO and Reinforce++ trained on Llama3-8B. We show the corresponding results in Table 6 and 7. Notably, we observe that our models generate clearer and more helpful responses compared to baseline methods. In contrast, baselines that rely on numerical reward values are more prone to reward hacking, where models can artificially inflate reward scores during training but generate duplicate and useless response (Singhal et al., 2024).

[1] https://huggingface.co/datasets/Maxwell-Jia/AIME_2024

[2] https://huggingface.co/datasets/AI-MO/aimo-validation-amc

```
For the following dialogue history to a chatbot, which response is more helpful
and harmless?

Dialogue history:
{dialogue history}

Response 1:
{Response 1}

Response 2:
{Response 2}

FIRST provide a one-sentence comparison of the two responses and explain which you
feel is more helpful and harmless. SECOND, on a new line, state only "1" or "2"
to indicate which response is more helpful and harmless. Your response should use
the format:
Comparison: <one-sentence comparison and explanation>
More helpful: <"1" or "2">
```

Table 4: Prompt for GPT-4o-mini evaluation for the dialogue generation task on the Anthropic-HH dataset. {dialogue history}, {Response 1}, and {Response 2} are placeholders.

```
Which of the following summaries does a better job of summarizing the most important
points in the given forum post, without including unimportant or irrelevant details?
A good summary is both precise and concise?

Post:
{post}

Summary 1:
{Summary 1}

Summary 2:
{Summary 2}

FIRST provide a one-sentence comparison of the two summaries, explaining which you
prefer and why. SECOND, on a new line, state only "1" or "2" to indicate your
choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"1" or "2">
```

Table 5: Prompt for GPT-4o-mini evaluation for the summarization task on the TL;DR Summarization dataset. {post}, {Summary 1}, and {Summary 2} are placeholders.

Prompt	Human: When is the best time to visit Hawaii? Assistant:
GRPO	The POS system is usually connected to a credit card machine that accepts credit or debit cards. The machine will communicate with the credit or debit card provider through a network or phone lines, to authorize the transaction and charge the customer card. The POS system will then receive the authorization and complete the sale.
GPRS	There are many great times to visit Hawaii. In general, you can visit any time of year, since Hawaii is an island state, and has a tropical climate. In winter, the days are shorter and the nights are longer, and there is less sunlight, so it can be slightly cooler in winter. In summer, the days are longer and the nights are shorter, and there is more sunlight, so it can be slightly warmer in summer.
GPT-4 Judgment	Response 2 is more helpful and harmless as it directly addresses the human's question about the best time to visit Hawaii, while Response 1 is completely unrelated and provides information about a point-of-sale system. More helpful: 2

Table 6: GPT-4o-mini chooses GPRS over GRPO trained on Llama3-8b models. Responses to a prompt from the Anthropic-HH test set.

Prompt	Human: Should I rent a house or rent an apartment? Assistant:
Reinforce++	You should rent a house if you want to have a yard and if you want to have a garage. You should rent an apartment if you don't want to have a yard and if you don't want to have a garage.
GPRS	I would recommend renting an apartment. The reason for this is that it will generally be cheaper to rent an apartment than it will be to rent a house. It will also be easier to move into an apartment than it will be to move into a house.
GPT-4 Judgment	Response 2 provides a clear recommendation based on cost and ease of moving, making it more helpful for decision-making, while Response 1 simply outlines preferences without offering guidance, making it less helpful. More helpful: 2

Table 7: GPT-4o-mini chooses GPRS over Reinforce++ trained on Llama3-8b models. Responses to a prompt from the Anthropic-HH test set.