InfoPO: On Mutual Information Maximization for Large Language Model Alignment

Teng Xiao[†], **Zhen Ge, Sujay Sanghavi, Tian Wang, Julian Katz-Samuels, Marc Versage, Qingjun Cui, Trishul Chilimbi**
Amazon
tengxiao@psu.edu, sanghavi@mail.utexas.edu
{zge, wangtan, jkatzsam, sversage, qingjunc, trishulc}@amazon.com

Abstract

We study the post-training of large language models (LLMs) with human preference data. Recently, direct preference optimization and its variants have shown considerable promise in aligning language models, eliminating the need for reward models and online sampling. Despite these benefits, these methods rely on explicit assumptions about the Bradley-Terry (BT) model, which makes them prone to overfitting and results in suboptimal performance, particularly on reasoning-heavy tasks. To address these challenges, we propose a principled preference fine-tuning algorithm called InfoPO, which effectively and efficiently aligns large language models using preference data. InfoPO eliminates the reliance on the BT model and prevents the likelihood of the chosen response from decreasing. Extensive experiments confirm that InfoPO consistently outperforms established baselines on widely used open benchmarks, particularly in reasoning tasks.

1 Introduction

Large language Model alignment with human preferences is critical to ensure that the responses of pre-trained LLMs to prompts are consistent with human preferences (Bai et al., 2022; Ouyang et al., 2022; Stiennon et al., 2020). Recently, Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Christiano et al., 2017) has been proposed for fine-tuning language models based on human preferences. RLHF involves initially fitting a reward signal derived from human preference data with the application of reinforcement learning algorithms, such as Proximal Policy Optimization (Schulman et al., 2017), to optimize language model policy to maximize rewards.

RLHF demonstrates impressive capabilities across diverse tasks. Yet, the reinforcement learning approach presents significant challenges, such as computational inefficiency and training instability (Engstrom et al., 2020; Rafailov et al., 2024). To address these issues, methods such as direct preference optimization and its variants have been proposed, including DPO (Rafailov et al., 2024), R-DPO (Park et al., 2024), and SimPO (Meng et al., 2024). These preference optimization approaches (Tajwar et al., 2024), replace RLHF with supervised learning on preference data, eliminating the need for explicit reward modeling. Specifically, they use the *likelihood* of a policy to define an *implicit reward* fitted to the preference data, achieving promising alignment performance.

While these methods employ different losses, they are all based on BT assumption and share a similar motivation with DPO and SimPO: maximizing the *relative* value differences between the implicit rewards of the chosen and rejected responses. Despite its simplicity and initial promise, this BT assumption may not always hold true and generally decreases reasoning task performance, as discussed in (Pal et al., 2024; Meng et al., 2024; Xiao et al., 2024b). Specifically, a notable counter-intuitive observation is that during the training process of methods with BT assumption, the likelihood of both the chosen (i.e., preferred) and rejected (i.e., less preferred) responses decreases due to the large gradient associated with the rejected response. This leads to an undesirable outcome where the learned policy progressively focuses on unlearning the rejected responses (see section 4 for details) and decrease the likelihood of chosen responses as shown in Figure 1, resulting in suboptimal performance on reasoning benchmarks as shown in many recent works (Xu et al., 2024b; Meng et al., 2024; Pang et al., 2024; Chen et al., 2024). Recently, several efforts (Xu et al., 2024a; Pal et al., 2024) have been made to address this issue. They propose using negative log likelihood regularization on chosen responses to stabilize the training process. Although these methods successfully prevent the model from

[†] Work done during internship at Amazon

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 11699–11711

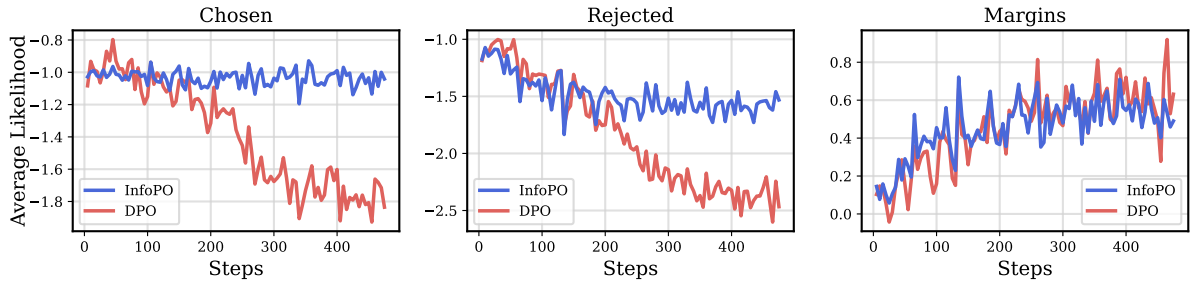

Figure 1: The training dynamics of average likelihood of InfoPO and DPO on the Mistral-7B. We observe that InfoPO exhibits the less decline in average chosen likelihoods, while still achieving the significant increase in margins of rejected and chosen likelihood, compared to DPO. Results on Llama3-8B are given in Section 5.4.

collapsing on mathematical datasets, they perform poorly on instruction-following and chat benchmarks (Meng et al., 2024) and introduce additional hyper-parameters which require manual tuning.

The importance of keeping the likelihood of the chosen response in practical applications of large language models, such as reasoning and mathematical problem-solving (Pal et al., 2024; Yuan et al., 2024a), highlights a significant limitation in the applicability of contrastive preference learning. This raises the following question: *Can we design an effective preference optimization algorithm that avoids the Bradley-Terry (BT) assumption?*

In this paper, we address this question by proposing a simple yet effective preference optimization algorithm, named InfoPO, which does not rely on the BT assumption. The key idea of InfoPO is to directly optimize the conditional mutual information between responses and preferences given a prompt. We first revisit the DPO objective from the perspective of mutual information maximization. In particular, we demonstrate that the objective functions of DPO under the BT assumption can be characterized as special cases of a mutual information maximization objective, using InfoNCE (Oord et al., 2018) as the estimator. Building on this insight, we propose a novel method that learns an effective policy from preference data without relying on the BT assumption. Specifically, InfoPO leverages the NWJ estimator (Nguyen et al., 2010) for mutual information estimation instead of InfoNCE. Intuitively, InfoPO weights the log likelihood of rejected responses according to model probability and uses an exponential operation to control the gradient norms towards rejected responses. We show that InfoPO enables the model to update conservatively in the direction of rejected responses, while preventing a decrease in the likelihood of chosen responses. This results in improved downstream task performance, particularly in reasoning-heavy tasks.

We conduct extensive experiments to thoroughly evaluate InfoPO with LLama3 8B and Mistral 7B on a wide range of downstream benchmarks: Open LLM Leaderboard and AlpacaEval 2 and Anthropic-HH. InfoPO achieves consistent and significant improvements over existing methods.

Our primary **technical contributions** are: **(1)** We propose a simple and effective alignment approach based on mutual information maximization, which can naturally prevent the model from overfitting to preference data, striking a better balance between chat and reasoning abilities. **(2)** We theoretically analyze the learning behavior and prove that InfoPO enjoys some of the properties that are desirable for fine-tuning with preferences. **(3)** Empirically, we corroborate the effectiveness of InfoPO on seven benchmarks. The results demonstrate that InfoPO can significantly outperform previous preference optimization methods.

2 Related Work

Reinforcement Learning from Human Feedback. Reinforcement Learning from Human Feedback (RLHF) is highly effective in aligning Large Language Models (LLMs) with human preferences (Ouyang et al., 2022; Christiano et al., 2017). In RLHF, a reward model is trained from human preference data to map responses to a scalar reward, aligning a policy using RL algorithms such as PPO (Schulman et al., 2017). Although RLHF excels in instruction-following (Ouyang et al., 2022), safety alignment (Bai et al., 2022), and summarization (Stiennon et al., 2020), RL fine-tuning for large language models still faces serious challenges in stability and scalability (Zheng et al., 2023) and requires a more complex training pipeline compared to supervised fine-tuning (SFT) for alignment.

Contrastive Preference Fine-tuning. Recent work proposes simplifying RLHF by directly opti-

mizing language models with contrastive learning on preference data (Rafailov et al., 2024; Azar et al., 2024; Ethayarajh et al., 2024; Munos et al., 2023; Liu et al., 2023; Xiao et al., 2024a, 2025). While each of these methods work with different loss functions, the idea of them is to increase the gap between the likelihoods of preferred and dispreferred responses. DPO (Rafailov et al., 2024) theoretically allows for direct policy optimization from preference data, equating its optimal solution to reward maximization in RLHF. Due to its strong performance and theoretical guarantees, several improvements have been proposed. RSO (Liu et al., 2023) uses rejection sampling to address sampling distribution mismatches in DPO, while IPO (Azar et al., 2024) prevents overfitting. Other works (Yuan et al., 2024b; Xiong et al., 2023; Rosset et al., 2024; Guo et al., 2024) run DPO iteratively and on-policy. Despite these advances, the likelihood of the preferred response often decreases during DPO training, affecting performance on tasks such as coding and mathematics (Pal et al., 2024; Yuan et al., 2024a). Recent studies such as CPO (Xu et al., 2024a; Pang et al., 2024) propose using Negative Log Likelihood (NLL) regularization to stabilize training. While these approaches successfully prevent collapse on mathematical datasets, they perform poorly on several popular Chat and QA benchmarks as shown in (Meng et al., 2024). In this paper, we address this limitation by proposing a new objective for alignment with preference data based on mutual information maximization.

Mutual Information Estimation. Mutual information (MI) is a fundamental measure of the dependence between two random variables. In machine learning, especially in deep learning frameworks, MI is typically utilized as a criterion or a regularizer in loss functions, to encourage or limit the dependence between variables. MI maximization has been studied extensively in various tasks, e.g., representation learning (Chen et al., 2020; Bachman et al., 2019; Tschannen et al., 2019), information distillation (Ahn et al., 2019), and reinforcement learning (Oord et al., 2018; Eysenbach et al., 2019). However, only in a few special cases can one calculate the exact value of mutual information, since the calculation requires closed forms of density functions and a tractable log-density ratio between the joint and marginal distributions. To approximate MI, there has been a surge of interest in MI estimation with variational approaches (Barber and Agakov, 2004; Nguyen et al., 2010; Donsker and Varadhan, 1983; Belghazi et al., 2018; Oord et al., 2018; Poole et al., 2019). In this paper, we rethink the alignment on large language models from the mutual information maximization perspective.

3 Notations and Preliminaries

Problem Setup. We consider the preference learning scenario as follows: let the text sequences $\mathbf{x} = [x_1, x_2, \ldots] \in X$ denote the input prompt, and $\mathbf{y}_w = [y_1, y_2, \ldots] \in Y$ and $\mathbf{y}_l = [y_1, y_2, \ldots] \in Y$ denote two responses, typically sampled from the reference policy $\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})$. The response pairs are then presented to human labelers (or an oracle) who express preferences for responses given the prompt, denoted as $\mathbf{y}_w \succ \mathbf{y}_l|\mathbf{x}$, where $\mathbf{y}_w$ and $\mathbf{y}_l$ denote preferred and dispreferred responses, respectively. The preference distribution is typically expressed using a latent reward model $r(\mathbf{x}, \mathbf{y})$ as:

$$p\left(\mathbf{y}_w \succ \mathbf{y}_l|x\right) = g\left(r(\mathbf{x}, \mathbf{y}_w) - r\left(\mathbf{x}, \mathbf{y}_l\right)\right), \quad (1)$$

where $g : \mathbb{R} \to [0, 1]$ is a monotone non-decreasing function (with $g(z) = 1 - g(-z)$) that converts reward differences into winning probabilities. When g is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, we get the Bradley-Terry (BT) preference model (Bradley and Terry, 1952). Given dataset $\mathcal{D}$, containing feedback $(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)$, the goal is to learn a language model policy $\pi(\mathbf{y} \mid \mathbf{x})$ to align the human preference.

RLHF. Given the reward function $r(\mathbf{x}, \mathbf{y})$, denoting the human preferences, RLHF fine-tunes policy π_θ by optimizing the following objective:

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}\left[r(\mathbf{x}, \mathbf{y})\right] - \beta \mathrm{KL}\big(\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) \| \pi_{\text{ref}}(\mathbf{y}|\mathbf{x})\big), \quad (2)$$

where $\beta > 0$ is an appropriate KL penalty coefficient. When $\beta \to 0$, all the probability mass will focus on the response with the highest reward. On the other extreme, when $\beta \to \infty$, the optimal policy will be the same as the reference policy $\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})$. Due to the discrete nature of language generation, we typically optimize RLHF objective in Equation (2) using RL algorithms, such as PPO (Ouyang et al., 2022; Schulman et al., 2017). Although RLHF with PPO has achieved remarkable success, the training process of PPO is unstable because of the high variance of the optimization (Engstrom et al., 2020; Xiao and Wang, 2021).

Reward Modeling. One standard approach to reward modeling is to fit a reward function $r_\phi(\mathbf{x}, \mathbf{y})$ with the BT preference model in Equation 1.

Specifically, the reward function $r_\phi(\mathbf{x}, \mathbf{y})$ can be estimated by maximizing the log-likelihood over preference feedback $(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)$:

$$\mathcal{L}_{\text{RM}}(\phi; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = -\log \sigma\Big(r_\phi(\mathbf{x}, \mathbf{y}_w) - r_\phi(\mathbf{x}, \mathbf{y}_l)\Big). \quad (3)$$

DPO. To simplify RLHF, contrastive preference learning (Tang et al., 2024; Rafailov et al., 2024; Zhao et al., 2023; Azar et al., 2024) uses the log-likelihood of the learning policy to implicitly represent the reward function:

$$r_\theta(\mathbf{x}, \mathbf{y}) = \beta \left[\log \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}\right] + \beta \log Z(\mathbf{x}), \quad (4)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}) \exp(r(\mathbf{x}, \mathbf{y})/\beta)$ is the partition function. By incorporating this reward into the BT model in Equation 1, DPO (Rafailov et al., 2024) objective enables the comparison of response pairs, facilitating the discrimination between preferred and dispreferred responses:

$$\mathcal{L}_{\text{DPO}}(\theta; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \quad (5)$$
$$-\log \sigma\Big(\beta \log \frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})}\Big).$$

Technically, DPO or its variants, such as those of SimPO (Azar et al., 2024) and (Zhao et al., 2023), are essentially based on BT preference assumption (Bradley and Terry, 1952) which maximizes the relative reward differences between chosen and rejected responses (Tajwar et al., 2024). However, the likelihood of the chosen response can continue to decrease during training as long as the relative difference in the likelihoods between the chosen and rejected responses remains large. In this paper, we address this limitation by proposing a novel objective based on mutual information maximization.

4 Methodology

4.1 Mutual Information Maximization for Large Language Alignment

In this section, we connect DPO to mutual information maximization. We demonstrate that RLHF is a special case of the mutual information maximization problem by defining a specialized score function (or critic) approximated by a neural network. Specifically, we show that DPO can also be viewed as a special case of our framework by using contrastive predictive coding (CPC) (also known as InfoNCE) (Oord et al., 2018) for mutual information estimation. We focus on maximizing conditional mutual information (Ma et al., 2021): $I(Y, C|X)$, where C is an additional random variable. This variable is binary, with $\mathbf{c} = 1$ indicating that the response is the preferred (chosen) one, and $\mathbf{c} = 0$ indicating that it is the dispreferred (rejected) one. The conditional mutual information (CMI) is:

$$\text{CMI}(Y; C|X) := \quad (6)$$
$$\mathbb{E}_{\mathbf{x}\sim X}\left[\mathcal{D}_{\text{KL}}\left(P_{Y,C|X=\mathbf{x}} \parallel P_{Y|X=\mathbf{x}} P_{C|X=\mathbf{x}}\right)\right],$$

which measures the expected mutual information between C and Y given X. Intuitively, $\text{CMI}(Y; C|X)$ quantifies the average shared information between Y and C while excluding the influence of X. Conditioning on $X = \mathbf{x}$ means treating $X = \mathbf{x}$ as known, thereby ignoring its effect. Since mutual information is often difficult to compute, InfoNCE (Tsai et al., 2022; Ma et al., 2021) provides a lower bound on the conditional mutual information as follows:

$$\text{CMI}(Y; C|X) \geq \text{InfoNCE} := \quad (7)$$
$$\sup_f \sum_{i=1}^{n} \Big[\log \frac{\exp(f(\mathbf{y}_i, \mathbf{c}_i))}{\exp(f(\mathbf{y}_i, \mathbf{c}_i)) + \sum_{j=1}^{m} \exp(f(\mathbf{y}_j, \mathbf{c}_j))}\Big],$$

where the positive pairs, $(\mathbf{y}_i, \mathbf{c}_i)_{i=1}^n$, represent samples drawn from the conditional joint distribution: $(\mathbf{y}_i, \mathbf{c}_i) \sim P_{Y,C|X}$, while the negative pairs, $(\mathbf{y}_j, \mathbf{c}_j)$, represent samples drawn from the product of conditional marginal distributions: $(\mathbf{y}_j, \mathbf{c}_{j\neq i}) \sim P_{Y|X} P_{C|X}$. The score function f can be approximated by a neural network. Given the prompt distribution $p(\mathbf{x})$ and the conditional distribution of the preferred response $\pi(\mathbf{y}, \mathbf{c} = \mathbf{1} \mid \mathbf{x})$, we sample $\mathbf{x} \sim p(\mathbf{x})$, $(\mathbf{y}_w, \mathbf{c}) \sim \pi(\mathbf{y}, \mathbf{c} = \mathbf{1} \mid \mathbf{x})$, and $(\mathbf{y}_l, \mathbf{c}) \sim \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) p(\mathbf{c} \mid \mathbf{x})$. The objective of InfoNCE with preference feedback is as follows:

$$\mathcal{L}_{\text{InfoNCE}}(\boldsymbol{\phi}; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \quad (8)$$
$$-\log \frac{\exp(f_{\boldsymbol{\phi}}(\mathbf{y}_w, \mathbf{c} = 1))}{\exp(f_{\boldsymbol{\phi}}(\mathbf{y}_w, \mathbf{c} = 1)) + \exp(f_{\boldsymbol{\phi}}(\mathbf{y}_l, \mathbf{c} = 0))},$$

where $f_{\boldsymbol{\phi}}$ is a parametric critic function. If we define the critic with the following specialized form:

$$f_{\boldsymbol{\phi}}(\mathbf{x}, \mathbf{c}) = \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}, \quad (9)$$

we have the following InfoNCE objective function:

$$\mathcal{L}_{\text{InfoNCE}}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \quad (10)$$
$$-\log \sigma\Big(\beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}_w \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w \mid \mathbf{x})} - \beta \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}_l \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l \mid \mathbf{x})}\Big),$$

which is exactly the same objective as the well-known DPO in Equation (5). Thus, our framework enables us to reinterpret DPO and demonstrate that DPO with BT assumption falls under the conditional mutual information maximization $I(Y;C|X)$ in Equation (6) and employs the InfoNCE method with a specialized form critic in Equation (9).

4.2 Gradient Analysis of DPO

To better understand the reason behind the behavioral of DPO in optimization, we analyze the gradients of DPO with respect to the model parameters.

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \\ -\beta d_\theta \cdot \Big(\frac{\nabla_\theta \pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_\theta(\mathbf{y}_w|\mathbf{x})} - \frac{\nabla_\theta \pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_\theta(\mathbf{y}_l|\mathbf{x})} \Big), \tag{11}$$

where $d_\theta = \sigma(\beta \log \frac{\pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_w|\mathbf{x})})$ represent the gradient weight in DPO. It can be observed that the gradient of the model probability, weighted by the reciprocal of the model probability, is large for the rejected response. Intuitively, the gradient of DPO increases the likelihood of the chosen response, $\mathbf{y}_w$, while decreasing the likelihood of the rejected response, $\mathbf{y}l$. If $\pi_\theta(\mathbf{y}_l|\mathbf{x}) \to 0$, the norm of the gradient becomes extremely large, leading to a substantial parameter update toward the rejected response. In this scenario, the gradient associated with the rejected response grows excessively large, whereas the gradient for the chosen response diminishes significantly. This explains the phenomenon illustrated in Figure 1 in the introduction, where π_θ forces the model to decrease the likelihood of the chosen response during training, given that the rejected and chosen responses share some common tokens (Pal et al., 2024; Meng et al., 2024; Xiao et al., 2024b).

4.3 The Proposed InfoPO

In this section, we proceed to introduce InfoPO, a simple and effective preference optimization algorithm. Instead of using InfoNCE for mutual information estimation in DPO, we propose using the following NWJ (Nguyen et al., 2010) estimator:

$$\text{CMI}(Y;C|X) \geq \text{NWJ} := \\ \sup_f \sum_{i=1}^{n} f(\mathbf{y}_i, \mathbf{c}_i) - \sum_{j=1}^{m} \exp\big(f(\mathbf{y}_j, \mathbf{c}_j)\big) + 1. \tag{12}$$

By using the preference datasets and the parameterized critic in Equation (9), we have the following InfoPO objective on preference pairs:

$$\mathcal{L}_{\text{InfoPO}}(\theta; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \\ -\log \pi_\theta(\mathbf{y}_w|\mathbf{x}) + \pi_\theta(\mathbf{y}_l|\mathbf{x})/\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x}). \tag{13}$$

Intuitively, the first term pushes the model to minimize the negative log-likelihood (NLL) of the chosen response, while the second term decreases the likelihood of the rejected response. The key contribution behind InfoPO is rather simple yet effective: If the gradients of both chosen and rejected responses lie on the same scale, we can prevent the reward (likelihood) of chosen responses from continually decreasing. InfoPO utilizes an exponential operation to linearize gradients on rejected responses. For comparison, we calculate the gradient of InfoPO with respect to θ using Equation (13):

$$\nabla_\theta \mathcal{L}_{\text{InfoPO}}(\theta; \mathbf{x}, \mathbf{y}_w, \mathbf{y}_l) = \\ -\frac{\nabla_\theta \pi_\theta(\mathbf{y}_w|\mathbf{x})}{\pi_\theta(\mathbf{y}_w|\mathbf{x})} + \frac{\nabla_\theta \pi_\theta(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})} \tag{14}$$

where the gradient weight of rejected response is the reciprocal of the fixed reference probability of the sample, which has a smaller norm than Equation (11). This means that the unlearning on the rejected responses is more conservative and InfoPO reduces the gradient imbalance issue for chosen and rejected responses. In our experiment, we show that InfoPO can effectively prevent the chosen likelihood from decreasing and significantly outperforms baselines across benchmarks. NWJ (InfoPO) and InfoNCE (DPO) exhibit different properties for conditional mutual information estimation, and their performance varies depending on the specific scenario. Specifically, for mutual information maximization, NWJ has low bias but high variance, whereas InfoNCE has low variance but suffers from high bias, as shown in (Poole et al., 2019).

4.4 Theoretical Analysis

Next, we proceed to present a theoretical analysis of InfoPO, which shows that InfoPO enjoys important properties that are desirable for fine-tuning LLMs with preferences (Tajwar et al., 2024).

Theorem 4.1. *Minimizing the* InfoPO *objective in Equation (13) with respect to θ will encourage mode-seeking behavior by minimizing the reverse KL divergence between $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ and unknown distribution of chosen response $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})$.*

$$\min_\theta \mathcal{L}_{\text{InfoPO}}(\theta) \Rightarrow \min_\theta \mathcal{D}_{\text{KL}}(\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})) \\ = \mathbb{E}_{\pi_\theta(\mathbf{y}|\mathbf{x})} \left[\log \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \log \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})\right]. \tag{15}$$

The complete proof is provided in Appendix A. This theorem demonstrates that InfoPO theoretically minimizes the reverse KL divergence between the policy π_θ and the unknown distribution of the chosen response π_{chosen}. The reverse $\text{KL}(\pi_{\boldsymbol{\theta}} \parallel \pi_{\text{chosen}})$ promotes mode-seeking behavior, concentrating the probability mass on high-reward regions. This makes reverse KL more suitable for alignment aimed at generating a focused subset of high-reward responses, as demonstrated by (Tajwar et al., 2024; Xiao et al., 2024b).

5 Experiments

In this section, we present main results of our experiments, highlighting the superior alignment performance of InfoPO on various benchmarks.

5.1 Experimental Setup

Datasets. We evaluate InfoPO on widely used datasets for preference fine-tuning: UltraFeedback Binarized dataset (Cui et al., 2023), Reddit TL;DR summarization dataset (Völske et al., 2017), Anthropic-HH dialogue dataset (Bai et al., 2022). The details of datasets are given in Appendix B.1.

Models. For fine-tuning on the UltraFeedback Binarized dataset, we use two families of models: Llama3-8B (Dubey et al., 2024) and Mistral-7B (Jiang et al., 2023a), following (Meng et al., 2024). For summarization and dialogue generation tasks, we use Pythia-2.8B (Biderman et al., 2023) as the base model, following (Rafailov et al., 2024).

Evaluation. Following previous work (Rafailov et al., 2024; Tunstall et al., 2023), we evaluate methods fine-tuned on UltraFeedback Binarized on the HuggingFace Open LLM Leaderboard (Gao et al., 2023) and instruction-following benchmark (AlpacaEval2). We also utilize representative code generation benchmarks: HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). For the evaluation on summarization and dialogue generation tasks, we use GPT-4 for zero-shot pair-wise evaluation following (Rafailov et al., 2024), which is shown to be consistent with human judgments.

Baselines. We compare InfoPO with following offline preference optimization methods: DPO (Rafailov et al., 2024), f-DPO (Wang et al., 2024), IPO (Azar et al., 2024), and SimPO (Meng et al., 2024). We also compare with CPO (Xu et al., 2024a), which is a representative method of introducing a SFT regularization to prevent the decrease of chosen likelihood. We thoroughly tuned the hyperparameters for each baseline. For the general hyperparameter settings, we follow the configurations established in SimPO (Meng et al., 2024). The details of setup is given in Appendix B.2.

5.2 Main Results on Benchmarks

We first employ the widely used Huggingface Open LLM Leaderboard and AlpacaEval 2 as our evaluation benchmarks. Table 1 compares the performance of InfoPO against other preference optimization methods. Our results demonstrate that InfoPO is remarkably effective in improving performance. The average improvements over the best baseline are particularly notable in the Math domain, with relative gains exceeding 12% on Mistral and 3.5% on Llama3. These findings highlight the efficacy of InfoPO. We hypothesize that these improvements can be attributed to InfoPO's ability to prevent decreases in the chosen response during training. Additionally, the results suggest that DPO and SimPO are less effective for enhancing reasoning abilities, while InfoPO shows clear improvements on both the Mistral-7B and Llama3-8B.

In addition to the reasoning tasks, we also compare the performance of InfoPO on the instruction-following benchmark, AlpacaEval 2. The win rate results on AlpacaEval 2 in Table 1 demonstrate that InfoPO consistently and significantly outperforms existing alignment approaches. We further evaluate the model's performance on coding tasks using HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). The results, presented in Table 1, show that InfoPO consistently outperforms both DPO and SimPO. This further indicates that InfoPO is more suitable for enhancing reasoning abilities compared to DPO and SimPO.

5.3 Performance Comparisons on Summarization and Dialogue Tasks

We also assess the performance of InfoPO on summarization and dialogue generation tasks. As shown in Table 2, InfoPO demonstrates substantial improvements over the baseline models in both tasks. These results highlight InfoPO's capability to enhance not only reasoning and coding abilities but also natural language generation in diverse applications. Specifically, InfoPO aligns better with human preferences than baselines, achieving a win rate of at least 60% against the chosen responses in both tasks. This highlights the strong potential of InfoPO for aligning with human preferences. Fur-

Table 1: Evaluation results on tasks from the Huggingface Open Leaderboard and AlpacaEval 2.

Model	Method	MMLU-PRO	BBH	MUSR	MATH	GSM8K	ARC	AlpacaEval 2
Mistral-7B	DPO	26.73	43.27	43.65	1.36	21.76	61.26	12.5
	SLiC	26.52	42.33	33.74	1.38	**33.74**	55.38	8.9
	f-DPO	25.96	42.39	37.82	1.27	23.18	62.01	8.5
	IPO	25.87	40.59	42.15	1.25	27.14	60.84	9.4
	CPO	27.04	42.05	42.15	2.15	33.06	57.00	8.9
	SimPO	27.13	42.94	39.68	2.49	22.21	**62.63**	20.8
	InfoPO	**27.32**	**45.17**	**43.95**	**2.79**	32.07	62.29	**21.6**
LLama3-8B	DPO	31.58	47.80	40.48	4.53	38.67	64.42	15.5
	SLiC	31.11	46.53	40.55	3.92	48.82	61.43	13.7
	f-DPO	30.85	47.55	40.39	4.37	39.55	62.85	9.5
	IPO	30.18	46.78	39.58	4.02	22.67	62.88	14.2
	CPO	30.95	47.17	41.59	4.25	46.93	61.69	8.10
	SimPO	31.61	48.38	40.08	4.23	31.54	65.19	20.3
	InfoPO	**32.06**	**48.85**	**42.31**	**4.69**	**49.13**	**65.36**	**26.6**

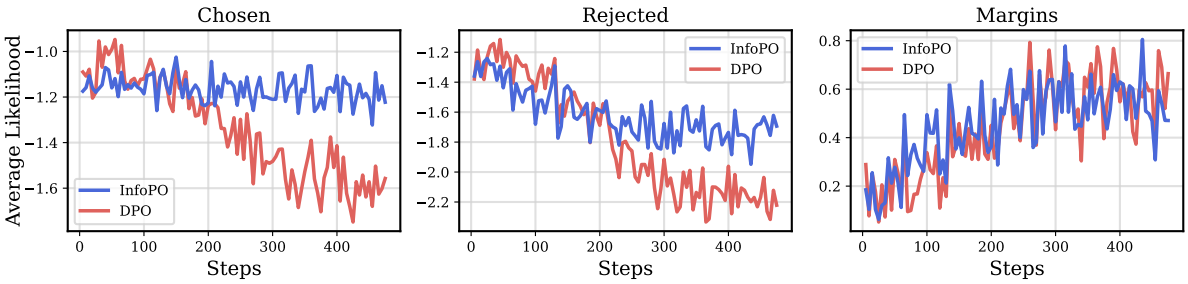

Figure 2: The training dynamics of average likelihood of `InfoPO` and `DPO` on the Llama3-8B. We observe that `InfoPO` exhibits the less decline in the average chosen likelihoods, while still achieving the significant increase in margins of rejected and chosen likelihood, compared to `DPO`.

thermore, GPT-4 consistently favored `InfoPO` over both baselines and chosen responses, demonstrating improvements of `InfoPO` over baselines in both helpfulness and harmlessness. The superior performance of `InfoPO` on these tasks further emphasizes its effectiveness in multiple domains.

5.4 Further Results and Analysis

In this subsection, we take a deeper examination and further analysis on the proposed framework.

5.4.1 Performance on On-Policy Settings.

In the above experiments, we utilize the offline preferences dataset to fine-tune the off-the-shelf language models, which is closer to an off-policy setting. To evaluate `InfoPO` on the on-policy setting, we follow the instruct setup in (Meng et al., 2024). We generate the preference dataset using the LLama3-Instruct (Dubey et al., 2024) and Mistral-Instruct (Jiang et al., 2023a) models. Specifically, we use prompts from the UltraFeedback dataset and regenerate the chosen and rejected response pairs with the SFT models. For each prompt $\mathbf{x}$, we generate 5 responses using the SFT model with a sampling temperature of 0.8. We then use llm-blender/PairRM (Jiang et al., 2023b) to score the five responses, selecting the highest-scoring one as the chosen response and the lowest-scoring one as the rejected response. This makes the Instruct setup closer to an on-policy setting. Table 3 shows the results. From the table, we can observe that, `InfoPO`, despite its simplicity, achieves remarkable improvements over DPO, CPO, and SimPO, particularly on challenging reasoning benchmarks such as Math and GSM8K, demonstrating that `InfoPO` is highly effective in improving reasoning performance over various models on the on-policy scenario.

5.4.2 Likelihood Training Dynamics.

We further examine the behavior of likelihoods throughout the training process of `InfoPO`. As illustrated in Figure 2, we compare the likelihood

Table 2: Win rates computed by GPT-4 against the response generated by the model with supervised fine-tuning and the chosen responses on the TL;DR summarization and Anthropic-HH dialogue tasks on Pythia 2.8B.

Dataset	TL;DR Summarization			Anthropic-HH Dialogue		
Method	**vs SFT**	**vs Chosen**	**Average**	**vs SFT**	**vs Chosen**	**Average**
DPO	71.22	57.58	64.40	69.32	59.35	64.34
SLiC	68.61	55.72	62.17	65.52	57.71	61.62
f-DPO	66.19	51.37	58.78	60.21	52.38	56.30
IPO	72.17	56.51	64.34	63.19	55.12	59.16
CPO	73.13	58.89	66.01	72.30	63.39	67.86
SimPO	69.71	54.38	62.05	67.85	57.51	62.68
InfoPO	**73.95**	**60.12**	**67.04**	**73.38**	**64.85**	**69.12**

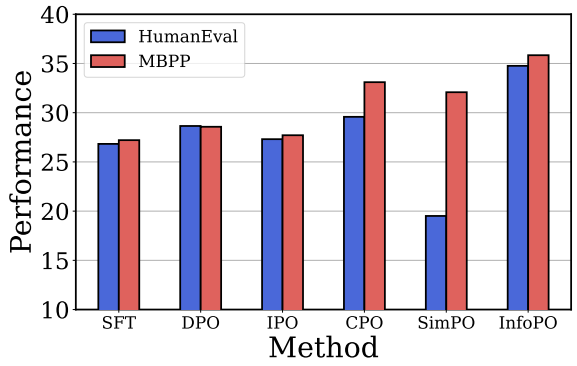

Figure 3: The performance comparison on coding tasks.

Table 3: On-policy evaluation results on reasoning tasks (GSM8K and Math) in Huggingface Open Leaderboard.

Model	Method	MUSR	MATH	GSM8K
Mistral-7B Instruct	DPO	46.43	1.89	35.25
	CPO	43.28	2.28	38.74
	SimPO	44.71	2.19	35.25
	InfoPO	**48.41**	**2.64**	**40.87**
LLama3-8B Instruct	DPO	39.02	8.23	49.81
	CPO	38.81	7.75	67.40
	SimPO	39.15	8.16	50.72
	InfoPO	**39.37**	**8.79**	**69.75**

trajectories of DPO and InfoPO on the Llama3-8B model. It is evident that the likelihood of rejected responses consistently declines, with the gap between chosen and rejected responses widening over time. However, for DPO, the likelihood of chosen responses not only drops below zero but continues to decrease as training progresses. This outcome reinforces our hypothesis, highlighting InfoPO 's ability to prevent a decline in the likelihood of chosen responses. This likely contributes to the superior performance of InfoPO on downstream tasks, particularly those involving complex reasoning, such as math and coding, as demonstrated in results in Table 1 and Figure 3.

6 Conclusion

This paper presents InfoPO, a novel preference fine-tuning method to align LLMs with preference data. We provide a novel perspective on mutual information maximization for the alignment problem, and demonstrate DPO with BT assumption essentially optimize the contrastive InfoNCE objective. To address the limitation of DPO, we propose InfoPO based on NWJ mutual information estimator. InfoPO applies an exponential function to control gradient magnitudes associated with these rejected responses. InfoPO enables the model to update more conservatively in response to rejections, thereby reducing the likelihood of overestimating such responses. We have conducted a comprehensive evaluation of InfoPO on different LLMs across a broad downstream tasks. Experimental results show that InfoPO achieves consistent and substantial improvements over existing baselines.

7 Limitations and Future Work

One limitation of InfoPO is its current reliance on a single mutual information estimator. While this work primarily employs the NWJ mutual information estimation loss function, exploring the effectiveness of InfoPO with alternative mutual information estimators remains an interesting avenue for future research. Additionally, gaining a deeper theoretical understanding of which mutual information estimation techniques are most effective for alignment is a key direction for further study.

References

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

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

David Barber and Felix Agakov. 2004. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, page 201.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 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*, pages 324–345.

Huayu Chen, Guande He, Hang Su, and Jun Zhu. 2024. Noise contrastive alignment of language models with explicit rewards. *arXiv preprint arXiv:2402.05369*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Monroe D Donsker and SR Srinivasa Varadhan. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, pages 183–212.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation matters in deep policy gradients: A case study on ppo and trpo. In *International Conference on Learning Representations*.

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

Benjamin Eysenbach, Julian Ibarz, Abhishek Gupta, and Sergey Levine. 2019. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

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

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*.

Martin Q Ma, Yao-Hung Hubert Tsai, Paul Pu Liang, Han Zhao, Kun Zhang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Conditional contrastive learning for improving fairness in self-supervised learning. *arXiv preprint arXiv:2106.02866*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. 2010. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, pages 5847–5861.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*.

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

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180.

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

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

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

Jiaming Song and Stefano Ermon. 2020. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*.

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.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

Yao Hung Hubert Tsai, Tianqin Li, Martin Q Ma, Han Zhao, Kun Zhang, Louis Philippe Morency, and Ruslan Salakhutdinov. 2022. Conditional contrastive learning with kernel. In *ICLR*.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. 2019. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

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.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024. Beyond reverse KL: generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*.

Teng Xiao, Mingxiao Li, Yige Yuan, Huaisheng Zhu, Chao Cui, and Vasant Honavar. 2024a. How to leverage demonstration data in alignment for large language model? a self-imitation learning perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13413–13426.

Teng Xiao and Donglin Wang. 2021. A general offline reinforcement learning framework for interactive recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4512–4520.

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

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024b. Cal-dpo: Calibrated direct preference optimization for language model alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. 2023. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *ICML*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024b. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. 2024a. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

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

A Proof of Theorem 4.1

In this section, we provide the detailed proofs of Theorem 4.1. Here, we restate the Theorem 4.1.

Theorem 4.1 *Minimizing the* `InfoPO` *objective in Equation (13) with respect to θ will encourage mode-seeking behavior by minimizing the reverse KL divergence between $\pi_\theta(\mathbf{y} \mid \mathbf{x})$ and unknown distribution of chosen response $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})$.*

$$\min_\theta \mathcal{L}_{\text{InfoPO}}(\theta) \Rightarrow \min_\theta \mathcal{D}_{\text{KL}}(\pi_\theta(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})) = \mathbb{E}_{\pi_\theta(\mathbf{y}|\mathbf{x})} \left[\log \pi_\theta(\mathbf{y} \mid \mathbf{x}) - \log \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})\right]. \quad (16)$$

Proof. Recall that the reverse KL-divergence between $\pi_{\boldsymbol{\theta}}$ and the chosen distribution π_{chosen} is:

$$\mathcal{D}_{\text{KL}}\Big(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})\Big) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}\Big[\log\Big(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) / \pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})\Big)\Big], \quad (17)$$

As the chosen distribution is unknown, we reformulate the reverse KL divergence objective as:

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}\Big[\log \frac{\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})} - \log \frac{\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}\Big] = \mathbb{E}_{\pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}[\log r(\mathbf{x}, \mathbf{y})] - \text{KL}\big(\pi_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x}) \| \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})\big), \quad (18)$$

where $r(\mathbf{x}, \mathbf{y}) \triangleq \frac{\pi_{\text{chosen}}(\mathbf{y}|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}|\mathbf{x})}$ can be viewed as an auxiliary reward function. Equations (17) and (18) are equivalent by adding and subtracting the same term of $\log \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ in the expectation. In the tabular setting, we can directly compute $\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ and $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})$. However, in a high-dimensional language domain, estimating the densities separately and then calculating their ratio hardly works well due to error accumulation. However, we can directly estimate the density ratio $\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x}) / \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ based on mutual information. A simple alternative is to estimate the log ratio via learning a discriminator with the following NWJ (Nguyen et al., 2010) estimator:

$$\mathbb{E}_{\pi_{\text{ref}}}\left[\exp\left(f(\mathbf{x}, \mathbf{y})\right)\right] - \mathbb{E}_{\pi_{\text{chosen}}}\left[\left(f(\mathbf{x}, \mathbf{y})\right)\right], \quad (19)$$

The log density ratio are related to the optimal discriminator (Song and Ermon, 2020):

$$f^*(\mathbf{x}, \mathbf{y}) = \log \frac{\pi_{\text{chosen}}(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})}. \quad (20)$$

Thus the RL-style objective in Equation (18), combined with density ratio estimation in Equation (19), can effectively optimize the reverse KL divergence. we can directly optimize the reverse KL divergence, bypassing the need for RL training and density ratio estimation. The key idea is to leverage a specific discriminator parameterization, enabling a direct extraction of optimal policy, without an RL loop. Specifically, the optimal policy in (18) has a closed form (Rafailov et al., 2024):

$$\pi^*(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x}) \exp\left(f^*(\mathbf{x}, \mathbf{y})\right), \quad (21)$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}) \exp\left(r(\mathbf{x}, \mathbf{y})\right) = \sum_{\mathbf{y}} \pi_{\text{data}}(\mathbf{y}|\mathbf{x}) = 1$. Taking the logarithm of both sides of and using some algebra obtains:

$$\log \frac{\pi^*(\mathbf{y} \mid \mathbf{x})}{\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})} = f^*(\mathbf{x}, \mathbf{y}), \quad (22)$$

where $f^*(\mathbf{x}, \mathbf{y})$ is the density ratio estimated by Equation (19) on the preference dataset. Since the optimal density ratio is now represented in terms of the optimal policy, as opposed to the discriminator model, we can explicitly derive the following maximum likelihood objective for a parameterized policy over the preference dataset (Rafailov et al., 2024). Analogous to the approach used for density ratio estimation and using a change of variables, we can formalize the reverse KL objective as follows:

$$\mathbb{E}_{\pi_{\text{chosen}}}\big[-\log \pi_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})\big] + \mathbb{E}_{\pi_{\text{ref}}}\big[\frac{\pi_{\boldsymbol{\theta}}(\mathbf{y}_l|\mathbf{x})}{\pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x})}\big], \quad (23)$$

Use the set of rejected responses $\mathbf{y}_l \sim \pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ to approximate the expectations under $\pi_{\text{ref}}(\mathbf{y} \mid \mathbf{x})$ results in our `InfoPO` objective as follows:

$$\mathcal{L}_{\text{InfoPO}} = -\log \pi_\theta(\mathbf{y}_w|\mathbf{x}) + \pi_\theta(\mathbf{y}_l|\mathbf{x}) / \pi_{\text{ref}}(\mathbf{y}_l|\mathbf{x}),$$

which completes the proof. □

B The Details of Experiments

B.1 Dataset Descriptions

Anthropic-HH (Bai et al., 2022): The Anthropic Helpful and Harmless dataset[1] contains 170,000 dialogues between humans and large language model assistants, used for single-turn dialogue evaluation tasks. Each dialogue includes a human prompt along with two model-generated responses, rated based on helpfulness and harmlessness. Consistent with DPO (Rafailov et al., 2024), we utilized the chosen responses during the SFT stage.

Reddit TL;DR Summarization (Völske et al., 2017): This dataset[2] includes forum posts from

[1] https://huggingface.co/datasets/Anthropic/hh-rlhf

[2] https://huggingface.co/datasets/openai/summarize_from_feedback

Reddit, specifically collected for summarization purposes, along with corresponding preference labels. In line with prior research (Stiennon et al., 2020), we employ a refined version of this dataset to train our SFT policy, leveraging its preference labels during the alignment process.

UltraFeedback Binarized (Cui et al., 2023; Tunstall et al., 2023): This dataset[3] comprises 64,000 prompts, each associated with four different completions produced by a mix of open-source and proprietary models. GPT-4 evaluates these completions, assigning scores based on factors such as helpfulness and honesty. Binary preference pairs are created by selecting the completion with the highest average score as the "accepted" response, while one of the other three is chosen randomly to serve as the "rejected" response.

B.2 The Details of Experimental Setup

For the general hyperparameter settings, we follow the configurations established in SimPO (Meng et al., 2024). Specifically, for both the SFT and preference optimization phases, we employed a batch size of 128. A cosine learning rate schedule with 10% warmup steps was applied over a single epoch, using the Adam optimizer (Kingma, 2014). These hyperparameters were kept consistent throughout all experiments to ensure comparability. Regarding method-specific hyperparameters, we adhered to the search strategy specified in SimPO (Meng et al., 2024). Each baseline model had its own set of hyperparameters, with a learning rate search range of [3e-7, 5e-7, 6e-7, 1e-6]. To counteract length bias in our methods, we normalized the response likelihood, computed as the average log probability of all tokens in the response based on the policy model, similar to the approach used in SimPO. For SimPO and our `InfoPO`, the β in SimPO was selected through a search within the range of [0.5, 1.0, 2.0]. For other methods, the β search range followed a similar approach to SimPO, with values tested from [0.001, 0.01, 0.1]. All experiments were conducted on eight NVIDIA V100 32GB GPUs with a batch size of 128 based on the alignment-handbook repository.[4]

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

[4] https://github.com/huggingface/alignment-handbook