Towards Efficient Exact Optimization of Language Model Alignment

Haozhe Ji [1] **Cheng Lu** [2] **Yilin Niu** [3] **Pei Ke** [1] **Hongning Wang** [1] **Jun Zhu** [2] **Jie Tang** [4] **Minlie Huang** [1]

Abstract

The alignment of language models with human preferences is vital for their application in real-world tasks. The problem is formulated as optimizing the model's policy to maximize the expected reward that reflects human preferences with minimal deviation from the initial policy. While considered as a straightforward solution, reinforcement learning (RL) suffers from high variance in policy updates, which impedes efficient policy improvement. Recently, direct preference optimization (DPO) was proposed to directly optimize the policy from preference data. However, we show that DPO derived based on the optimal solution of the problem leads to a compromised mean-seeking approximation of the optimal solution in practice. In this paper, we propose *efficient exact optimization* (EXO) of the alignment objective. EXO is guaranteed to optimize in the same direction as RL algorithms asymptotically for arbitrary policy parametrization. This leads to the same mode-seeking solution, while enables efficient optimization by circumventing the complexities of RL. We also compare our method to DPO with both theoretical and empirical analyses, and further demonstrate the advantages of our method over existing approaches on realistic human preference data. Code is available at `https://github.com/haozheji/exact-optimization`.

1. Introduction

Despite the proficiency of large language models, e.g., GPT-3 (Brown et al., 2020) in complex tasks under minimal supervision, they are still prone to produce harmful (Bai et al., 2022), biased (Bender et al., 2021), and unfaithful (Ji et al., 2023c) responses due to the heterogeneous sources of their pre-training corpora. Ensuring the large language models to generate desired responses that are in line with humans' ethical standards and quality preferences is crucial for the development of reliable AI systems.

The problem, well known as language model (LM) alignment with human preferences (Ouyang et al., 2022), is generally formulated as optimizing the LM policy π_θ to maximize the expected reward, which reflects human preferences regarding the completion $\boldsymbol{y}$ for a given prompt $\boldsymbol{x}$. The practical recipe is to train a reward model r_ϕ to predict the human-chosen response from a set of responses generated by an initial LM policy π_{init}. Yet, the challenge of acquiring substantial high-quality preference data often impedes accurate estimation of the ideal reward model. Consequently, this empirically learned reward model may lead to misspecified behaviors, particularly under the distributional shift between its training data and the data generated by π_θ (Gao et al., 2023). Therefore, the final objective of alignment additionally involves minimizing the reverse Kullback-Leibler (KL) divergence of π_θ from its initial distribution π_{init} with an intensity β, besides maximizing the expected reward:

$$\max_{\pi_\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\boldsymbol{y}\sim\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{init}}). \quad (1)$$

Due to the discrete nature of content generation from an LM (i.e., sampling $\boldsymbol{y}$ from $\pi_\theta(\cdot|\boldsymbol{x})$), the most widely adopted approach to optimize Eq. (1) is reinforcement learning (RL) (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). Albeit well studied even before the era of large language models, RL solutions are notorious for their poor stability due to the high variance in estimating the policy gradients or value functions, which potentially worsens sample complexity and thus compromises efficient convergence (Papini et al., 2018; Anschel et al., 2017).

As a remedy, direct preference optimization (DPO) was recently proposed to replace the RL solutions (Rafailov et al., 2023). Specifically, DPO defines a pair-wise preference loss on the estimated policy π_θ by leveraging the following policy-reward mapping in the optimal solution to Eq. (1):

$$\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_{\text{init}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta}r_\phi(\boldsymbol{x},\boldsymbol{y})} \quad (2)$$

[1]The Conversational AI (CoAI) Group, Tsinghua University [2]The Tsinghua Statistical Artificial Intelligence & Learning (TSAIL) Group, Tsinghua University [3]Zhipu AI [4]The Knowledge Engineering Group (KEG), Tsinghua University. Correspondence to: Haozhe Ji <jihaozhe@gmail.com>, Minlie Huang <aihuang@mail.tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

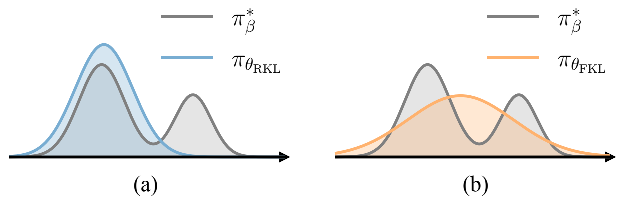


Figure 1. Illustration of different characteristics of (a) $\pi_{\theta_{\text{RKL}}}$ by minimizing the reverse KL (by EXO) and (b) $\pi_{\theta_{\text{FKL}}}$ by minimizing the forward KL (by DPO).

However, this optimal policy is obtained analytically while not considering the practical parametrization of π_θ. When π_θ is mis-specified and cannot perfectly capture the target π^*_β, our analysis demonstrates that DPO leads to a compromised approximation of π^*_β, which only covers the support of π^*_β while failing to capture its modes.

In this paper, we present an underexplored perspective of the alignment objective in Eq. (1): We prove that Eq. (1) is equivalent to *probability matching* between the parametrized policy π_θ and the optimal policy π^*_β defined in Eq. (2) measured by the reverse KL divergence $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi^*_\beta)$.

Based on the established equivalence, we propose *efficient exact optimization (EXO)* of the KL-regularized reward maximization objective. Specifically, we rigorously prove that irrespective of the policy's parametrization, EXO is guaranteed to improve π_θ in the same direction as the RL solutions asymptotically. In practice, we demonstrate that EXO facilitates efficient optimization towards this alignment objective with finite samples while bypassing the complexities of RL.

Under this probability matching perspective, we demonstrate that DPO actually corresponds to minimizing the forward KL divegence $\mathbb{D}_{\text{KL}}(\pi^*_\beta \| \pi_\theta)$. Though minimizing both the forward and reverse KL divergences lead to the same analytic solution, it is not necessarily achievable when taking into account the expressivity gap between the model families of π^*_β and π_θ (Lin et al., 2021). Under this realistic constraint, minimizing these two divergences converge to parametrized policies with different behaviors (Bishop & Nasrabadi, 2006). As illustrated in Figure 1, minimizing the reverse KL fosters a *mode-seeking* policy $\pi_{\theta_{\text{RKL}}}$ that concentrates to the principal modes of π^*_β (Chan et al., 2022), while minimizing the forward KL results in a *mean-seeking* policy $\pi_{\theta_{\text{FKL}}}$ that places large mass to the mean of different modes in π^*_β, which does not necessitate high probabilities under π^*_β. In the inference stage, $\pi_{\theta_{\text{RKL}}}$ is preferrably better than $\pi_{\theta_{\text{FKL}}}$ by capturing the main characteristics of π^*_β (Ji et al., 2023b).

We conduct a series of experiments to verify the effectiveness and scalability of EXO. We first systematically evaluate the efficiency of different approaches in trading off maximizing the oracle reward and minimizing the KL divergence during optimization of the alignment objective. Then, we conduct evaluations on the effectiveness of learning from real human preferences in various alignment benchmarks, involving summarization, dialogue generation, and instruction following tasks. Comprehensive empirical analyses substantiate our theoretical findings and demonstrate the advantageous performance of EXO over DPO and PPO.

Finally, we summarize our contributions in this paper:

- We reveal the underexplored equivalence between KL-regularizd reward maximization and minimizing the reverse KL divergence against the optimal policy for the language model alignment problem.
- We propose EXO, an algorithm towards *efficient exact* optimization of the KL-regularized reward maximization objective for alignment. Both theoretical and emprical results confirm its effectiveness.
- We show that DPO corresponds to minimizing the forward KL divergence, which is less effective in capturing the essential characteristics of the optimal policy.

2. Preliminaries

We first formally review the formulation and objective of the alignment problem. Then we review existing approaches that solve this problem via reinforcement learning and direct preference optimization, respectively.

2.1. Aligning Language Models with Human Preferences

Given a vocabulary $\mathcal{V}$, a language model defines a probability distribution $\pi(\boldsymbol{x}) = \prod_{t=1}^{n} \pi(x_t|x_1, \cdots, x_{t-1})$ over a sequence of tokens $\boldsymbol{x} = (x_1, \cdots, x_n)$. We apply π to a conditional generation task of interest with input space $\mathcal{X} = \mathcal{V}^m$ and output space $\mathcal{Y} = \mathcal{V}^n$ modeled by $\pi(\boldsymbol{y}|\boldsymbol{x}) = \pi(\boldsymbol{x}, \boldsymbol{y})/\pi(\boldsymbol{x})$.

The alignment procedure typically starts from supervised fine-tuning (SFT) the language model on a high-quality dataset $\mathcal{D}^{\text{sft}}$ via maximum likelihood estimation, which obtains the SFT policy π_{sft}.

Then a preference dataset $\mathcal{D}^{\text{pref}}$ is collected by asking humans to select a better response from $(\boldsymbol{y}_1, \boldsymbol{y}_2) \sim \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$ given a prompt $\boldsymbol{x}$ from the same domain of $\mathcal{D}^{\text{sft}}$. Let $\boldsymbol{y}_w$ and $\boldsymbol{y}_l$ be the chosen and rejected responses among $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ respectively according to human preferences.

A reward model $r_\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is usually learned on $\mathcal{D}^{\text{pref}}$ to act as a surrogate to expensive human labeling. The reward model is trained to prioritize $\boldsymbol{y}_w$ over $\boldsymbol{y}_l$ by minimizing the following pair-wise preference loss:

$$\mathcal{L}_r(r_\phi) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\text{pref}}}\left[-\log\frac{e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_w)}}{e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_w)}+e^{r_\phi(\boldsymbol{x},\boldsymbol{y}_l)}}\right].$$

Finally, a policy π_θ is learned to maximize the following alignment objective (Ziegler et al., 2019):

$$\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta\mathbb{D}_{\text{KL}}[\pi_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big). \quad (3)$$

Given a prompt $\boldsymbol{x}$ sampled from the dataset $\mathcal{D}^{\text{pref}}$, the objective seeks to find the π_θ that maximizes the expected reward while minimizes its reverse KL divergence against the SFT policy π_{sft} governed by the coefficient $\beta > 0$. The KL penalty keeps π_θ from moving too far from π_{sft} to avoid over optimization of the reward model.

The analytic solution that maximizes $\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta)$ takes the form of an energy-based model (EBM):

$$\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x}) = \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})\frac{e^{\frac{1}{\beta}r_\phi(\boldsymbol{x},\boldsymbol{y})}}{Z_\beta(\boldsymbol{x})}, \quad (4)$$

where $Z_\beta(\boldsymbol{x}) = \sum_{\boldsymbol{y}'\in\mathcal{Y}}\pi_{\text{sft}}(\boldsymbol{y}'|\boldsymbol{x})e^{\frac{1}{\beta}r_\phi(\boldsymbol{x},\boldsymbol{y}')}$ is the partition function. In Eq (4), the coefficient β can be considered as the temperature for controlling the strength of the reward model signal when sampling from $\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x})$.

2.2. RL Fine-Tuning

Due to the discrete nature of language generation, the objective in Eq. (3) is not differentiable with respect to π_θ, which prohibits supervised training. One standard approach is to use RL algorithms to optimize this objective. Ziegler et al. (2019) proposed to search for π_θ that maximizes a KL-regularized reward $r_\phi(\boldsymbol{x},\boldsymbol{y}) - \beta\log\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$. This can be achieved by policy gradient methods, such as Proximal Policy Optimization (PPO) (Schulman et al., 2017).

2.3. Direct Preference Optimization

To optimize π_θ directly using the preference data, Rafailov et al. (2023) rearranged Eq. (4) to express the reward function by the optimal policy π^*_β,

$$r_\phi(\boldsymbol{x},\boldsymbol{y}) = \beta\log\frac{\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} + \beta\log Z_\beta(\boldsymbol{x}). \quad (5)$$

Then they proposed to directly optimize the policy π_θ by replacing π^*_β with π_θ and substituting the corresponding reward function into a pair-wise preference loss:

$$\mathcal{L}_{\text{dpo}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim\mathcal{D}^{\text{pref}}}\Big[-\log\sigma\Big(\beta\log\frac{\pi_\theta(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta\log\frac{\pi_\theta(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_l|\boldsymbol{x})}\Big)\Big]. \quad (6)$$

3. Methodology

In this section, we start with a generalized alignment objective and present its equivalent form under the perspective of probability matching. Then we formally derive *efficient exact optimization* (EXO) of the generalized alignment objective while bypassing the necessity of employing any RL algorithms. Furthermore, we compare against DPO and demonstrate the advantage of our approach in terms of the distributional characteristics of the learned policy. All the proofs are provided in Appendix A.

3.1. From the Generalized Alignment Objective to Probability Matching

We first introduce a generalized alignment objective that distributes the intensity of the KL regularization regarding the SFT policy to both the parametrized policy π_θ and the reward model r_ϕ, which intuitively connects the regularization setting of DPO (Eq. (6)) that only regularizes π_θ and PPO (Eq. (3)) that only regularizes r_ϕ. In the following theorem, we present the formal definition and the property of the generalized alignment objective.

Theorem 3.1. *Let $\beta_\pi > 0, \beta_r > 0$ and $\beta_\pi\beta_r = \beta$. The generalized alignment objective is defined as*

$$\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi^{\beta_\pi}_\theta(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r\mathbb{D}_{\text{KL}}[\pi^{\beta_\pi}_\theta(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big), \quad (7)$$

where $\pi^{\beta_\pi}_\theta(\boldsymbol{y}|\boldsymbol{x})$ satisfies

$$\pi^{\beta_\pi}_\theta(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}. \quad (8)$$

Given unlimited model capacity, the optimal π_{θ^*} that maximizes $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$ satisfies $\pi_{\theta^*} = \pi^*_\beta$.

Intuitively, $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$ calculates the expectation of the reward regularized with the intensity β_r with respect to the policy regularized with the intensity β_π. As the total regularization intensity $\beta = \beta_r\beta_\pi$ is fixed, $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$ maintains the same analytic solution π^*_β as $\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta)$. Furthermore, it interpolates the policy regularization setting ($\beta_r = 1, \beta_\pi = \beta$) in the DPO objective[1] and the reward regularization setting ($\beta_r = \beta, \beta_\pi = 1$) in the PPO objective when continuously tunning β_r and β_π while keeping their product fixed. We also empirically show the effect of β_r and β_π beyond the impact on their product $\beta = \beta_r\beta_\pi$ in Appendix C.2.

Next, we derive an equivalent form of the generalized alignment objective by rearranging the elements in Eq. (7), which offers a new insight of the alignment problem from the probability matching perspective. The detailed derivation can be

[1] $\beta\log\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} = \log\frac{\pi^{\beta}_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} + C(\boldsymbol{x})$ where $C(\boldsymbol{x})$ depends only on $\boldsymbol{x}$, which does not effect the DPO objective in Eq. (6).

found in Appendix A.2.

$$\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) = -\beta_r \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\big[\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x}))\big] + \beta_r \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\big[\log Z_{\beta_r}(\boldsymbol{x})\big], \tag{9}$$

As the second term is a constant with respect to π_θ, Eq. (9) reveals that maximizing the generalized alignment objective $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ with respect to π_θ is equivalent to minimizing the expected reverse KL divergence $\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}\|\pi_{\beta_r}^*)$, where $\pi_\theta^{\beta_\pi}$ is a function of π_θ as defined in Eq. (8).

This equivalence implies the possibility of converting the reward maximization problem into a supervised divergence minimization problem, which is able to circumvent the poor stability and low sample efficiency issue caused by high variance in RL solutions (Papini et al., 2018; Anschel et al., 2017). In the following, we introduce our approach towards exact optimization of this generalized alignment objective by practically realizing the probability matching objective. Without loss of generality, our results remain valid for the original alignment objective in Eq. (3).

3.2. Efficient Exact Optimization of the Generalized Alignment Objective

We now formally derive EXO which optimizes the generalized alignment objective $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ by realizing the reverse KL divergence $\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}\|\pi_{\beta_r}^*)$. We start with the general setting of language model alignment which preassumes the existence of a reward model r_ϕ; then we consider the case of learning directly from the preference data.

To facilitate policy optimization with straightforward gradient back propagation, we rewrite $\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}\|\pi_{\beta_r}^*)$ under the expectation of the proposal policy π_{sft}:

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}\|\pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}\right].$$

The above expression can be further simplified by defining the log ratio as $f_\theta(\boldsymbol{x},\boldsymbol{y}) = \log\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) - \log\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$:

$$\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}\|\pi_{\beta_r}^*) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}\log\frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}}{\frac{1}{Z_{\beta_r}(\boldsymbol{x})}e^{\frac{r_\phi(\boldsymbol{x},\boldsymbol{y})}{\beta_r}}}\right].$$

As the intractable nature of $Z_{\beta_r}(\boldsymbol{x})$ hinders direct estimation via Monte Carlo simulation, we propose a practical way to estimate this term by first drawing multiple samples from π_{sft}, and then calculating the reverse KL between the probability distributions defined by $f_\theta(\boldsymbol{x},\boldsymbol{y})$ and $\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})$ over these samples via self-normalization respectively.

• **Learning from a reward model.** Formally, given $K>1$ i.i.d. completions $\boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1,\cdots,\boldsymbol{y}_K\}$ drawn from $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$, we define an empirical distribution p_{f_θ} by normalizing the exponential reward $e^{f_\theta(\boldsymbol{x},\boldsymbol{y})}$ over the K samples:

$$p_{f_\theta}(i|\boldsymbol{y}_{1:K},\boldsymbol{x}) = \frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^K e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}}. \tag{10}$$

Recall that $f_\theta(\boldsymbol{y}|\boldsymbol{x}) = \log\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) - \log\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$ and $\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}$, Eq. (10) can be rewritten into a form that explicitly depends on π_θ:

$$p_{f_\theta}(i|\boldsymbol{y}_{1:K},\boldsymbol{x}) = \frac{e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^K e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}}, \tag{11}$$

where β_π can be regarded as the inverse temperature that modulates the empirical distribution defined by the log ratio between π_θ and π_{sft}. Similarly, we define a distribution p_r over the K samples modeled by the reward model r_ϕ:

$$p_{r_\phi}(i|\boldsymbol{y}_{1:K},\boldsymbol{x}) = \frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^K e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}. \tag{12}$$

Finally, we translate the original objective of reward maximization $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ into the reverse KL between p_{f_θ} and p_{r_ϕ} over $\boldsymbol{y}_{1:K}$ sampled from π_{sft}:

$$\mathcal{L}_{\text{exo}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\Big[\mathbb{D}_{\text{KL}}\big(p_{f_\theta}(\cdot|\boldsymbol{y}_{1:K},\boldsymbol{x})\|p_{r_\phi}(\cdot|\boldsymbol{y}_{1:K},\boldsymbol{x})\big)\Big]. \tag{13}$$

The complete form of $\mathcal{L}_{\text{exo}}$ is presented in Eq. (23) in Appendix A.3. Besides its practical simplicity for implementation, we also elucidate its theoretical attributes by characterizing its connection with the generalized alignment objective in Theorem 3.2.

Theorem 3.2. *Following $\pi_{\beta_r}^*$, $\pi_\theta^{\beta_\pi}$ and $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ defined in Eq. (4), (8), and (7), for $K\to\infty$ and arbitary θ, the gradient of $\mathcal{L}_{\text{exo}}(\pi_\theta)$ satisfies*

$$\nabla_\theta\mathcal{L}_{\text{exo}}(\pi_\theta) = \nabla_\theta\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\big[\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x}))\big] \tag{14}$$

$$= -\frac{1}{\beta_r}\nabla_\theta\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}). \tag{15}$$

Theorem 3.2 suggests the optimization direction for π_θ during the intermediate optimization steps for minimizing $\mathcal{L}_{\text{exo}}(\pi_\theta)$ aligns with the direction required to maximize the generalized alignment objective $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ asymptotically, when sufficient sample population is provided. In §4.1, we show that EXO achieves encouraging convergence in practical scenarios with only a finite K. Again, as a special case, $\nabla_\theta\mathcal{L}_{\text{exo}}(\pi_\theta)$ aligns with $\nabla_\theta\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta)$ when $\beta_\pi=1, \beta_r=\beta$, which effectively connects with Eq. (3). The result in Theorem 3.2 is crucial, as it sheds light on exact optimization of the generalized alignment objective via a simple density matching approach with strong theoretical guarantees. In Appendix A.5, we provide a mechanistic understanding of the gradient $\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta)$ which is a weighted sum of the gradients $\nabla_\theta \log \pi_\theta(\boldsymbol{y}_k|\boldsymbol{x})$. The weight is proportional to $\mathbb{D}_{\text{KL}}(p_{f_\theta} \| p_{r_\phi})$ which characterizes the *distributional* gap biased by a *point-wise* correction on the sample $\boldsymbol{y}_k$.

• **Learning from human preference data.** In situations where only preference data is accessible, we devise an empirical formulation of $\mathcal{L}_{\text{exo}}$. Given a preference dataset $\mathcal{D}^{\text{pref}}$ where each prompt $\boldsymbol{x}$ is paired with $\boldsymbol{y}_w$ and $\boldsymbol{y}_l$ denoting the chosen and rejected completions. This binary supervision can be effectively transformed into an empirical distribution of p_{r_h} defined by the underlying reward r_h of human preference. To avoid infinity when calculating KL divergence, we smooth the one-hot distribution into a soft distribution, i.e., $p_{r_h}(w|\boldsymbol{y}_w, \boldsymbol{y}_l, \boldsymbol{x}) = 1 - \varepsilon$ and $p_{r_h}(l|\boldsymbol{y}_w, \boldsymbol{y}_l, \boldsymbol{x}) = \varepsilon$, where $\varepsilon > 0$ is a hyperparameter. p_{f_θ} can still be computed according to Eq. (11) over $\boldsymbol{y}_w$ and $\boldsymbol{y}_l$. As a result, we present the EXO objective on the prefence data by setting $K = 2$ and substituting r_ϕ with r_h in Eq. (13):

$$\mathcal{L}_{\text{exo-pref}}(\pi_\theta) = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{y}_l) \sim \mathcal{D}^{\text{pref}}} \Big[\mathbb{D}_{\text{KL}}\big(p_{f_\theta}(\cdot|\boldsymbol{y}_w, \boldsymbol{y}_l, \boldsymbol{x}) \| p_{r_h}(\cdot|\boldsymbol{y}_w, \boldsymbol{y}_l, \boldsymbol{x})\big)\Big]. \quad (16)$$

In practice, π_{sft} is fine-tuned on either $\mathcal{D}^{\text{sft}}$ which is collected from the same domain as $\mathcal{D}^{\text{pref}}$, or the chosen completions in $\mathcal{D}^{\text{pref}}$ when $\mathcal{D}^{\text{sft}}$ is not available. This closes the distributional gap between π_{sft} and the unknown distribution that generates the preference data $\mathcal{D}^{\text{pref}}$.

3.3. Comparing with DPO under the Perspective of Probability Matching

Under the perspective of probability matching, we formally demonstrate that the DPO objective corresponds to the *forward* KL which is essentially different from the reverse KL required by the alignment objective $\mathcal{J}^\beta_{\text{lhf}}(\pi_\theta)$ in Eq. (3). We then analyze their differences under realistic constraints of model capacities.

We first consider the general form of the DPO objective. Given K completions $\boldsymbol{y}_{1:K} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_K\}$ drawn from $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$ and a reward model r_ϕ that captures human prefence, we generalize $\mathcal{L}_{\text{dpo}}$ by substituting the sigmoid function with softmax over K responses and replacing the one-hot label with a soft distribution defined by r_ϕ:

$$\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}^{\text{pref}}} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})} \Bigg[-\sum_{i=1}^{K} \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x}, \boldsymbol{y}_i)}}{\sum_{j=1}^K e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x}, \boldsymbol{y}_j)}} \log \frac{e^{\beta_\pi \log \frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^K e^{\beta_\pi \log \frac{\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}} \Bigg]. \quad (17)$$

Upon substituting r_ϕ with r_h and setting $K = 2$, $\mathcal{L}_{\text{dpo-rw}}$ simplifies to $\mathcal{L}_{\text{dpo}}$. In the following, we build connection of $\mathcal{L}_{\text{dpo-rw}}(\pi_\theta)$ to the forward KL divegence $\mathbb{D}_{\text{KL}}(\pi^*_{\beta_r} \| \pi^{\beta_\pi}_\theta)$.

Theorem 3.3. *With $\pi^*_{\beta_r}$, $\pi^{\beta_\pi}_\theta$ and p_{r_ϕ} defined in Eq. (4), (8) and (12) respectively, for $K \to \infty$ and arbitary θ, the gradient of $\mathcal{L}_{\text{dpo-rw}}(\pi_\theta)$ satisfies*

$$\nabla_\theta \mathcal{L}_{\text{dpo-rw}}(\pi_\theta) = \nabla_\theta \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}^{\text{pref}}} \big[\mathbb{D}_{\text{KL}}(\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}) \| \pi^{\beta_\pi}_\theta(\boldsymbol{y}|\boldsymbol{x}))\big]. \quad (18)$$

In the following, we consider $\beta_\pi = 1, \beta_r = \beta$ to simplify the analysis, while the results still hold for general settings. Putting the results of Theorem 3.2 and Theorem 3.3 together, we readily connect $\mathcal{L}_{\text{exo}}$ and $\mathcal{L}_{\text{dpo-rw}}$ with two divergences, i.e., the reverse KL, $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi^*_\beta)$, and the forward KL, $\mathbb{D}_{\text{KL}}(\pi^*_\beta \| \pi_\theta)$, respectively. To provide a clear image, we illustrate the interrelationship among the objectives of EXO, DPO, and the objective of alignment in Figure 5, as presented in Appendix A.6. Although minimizing these two divergences leads to the same analytic solution π^*_β, they converge to two distinct solutions when considering the expressivity bottleneck of the practical model parametrization.

Specifically, the LM policy π_θ is commonly parametrized as an auto-regressive (AR) model, which enables efficient sampling due to the employment of local normalization. However, the optimal policy $\pi^*_\beta(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x}) \exp(\frac{1}{\beta} r_\phi(\boldsymbol{x}, \boldsymbol{y}))$ defined as an EBM trade-offs sampling efficiency with modeling capacity. Notably, Lin et al. (2021) rigorously proved that AR models cannot perfectly capture all possible distributions defined by EBMs in terms of supports or rankings based on the computational complexity theories. From an empirical view, this result is also intuitive because the reward model as a discriminator is more flexible in distribution modeling than the auto-regressive generator.

Under the practical constraint that π_θ cannot perfectly represent π^*_β, minimizing the forward and reverse KL results in two policies $\pi_{\theta_{\text{FKL}}}$ and $\pi_{\theta_{\text{RKL}}}$ with different properties. One well-known fact is that $\pi_{\theta_{\text{FKL}}}$ is *mean-seeking* while $\pi_{\theta_{\text{RKL}}}$ is *mode-seeking* (Bishop & Nasrabadi, 2006). In Figure 1, we consider an illustrative picture of fitting a unimodal π_θ to a multi-modal target π^*_β. To maintain a minimal forward KL divergence, $\pi_{\theta_{\text{FKL}}}$ must encompass all the modes of π^*_β (regions where π^*_β exhibits significant values). However due to the representational constraints, $\pi_{\theta_{\text{FKL}}}$ tends to overpresents the mean of different modes of π^*_β, potentially extending into the long tail region of π^*_β (Chan et al., 2022; Ji et al., 2023a). On the other hand, $\pi_{\theta_{\text{RKL}}}$ can select one mode of π^*_β without causing the reverse KL to explode, meaning that $\pi_{\theta_{\text{RKL}}}$ will effectively capture the major mode of π^*_β under realistic model capacity. In §4.1, we empirically demonstrate the results of optimizing these two divergences in practice.

Within the context of language model alignment, reverse KL is preferred for generating samples according to the *evaluation* of the optimal policy. Conversely, forward KL tradeoffs preference evaluation with sample diversity, which is rational only if the samples are valid under the evaluation. To some extent, the reverse KL can also effectively capture this rational diversity, as it maximizes the policy's entropy to prevent distributional collapse to a single point.

4. Experiments

We verify the effectiveness of EXO via extensive experimentations. In §4.1, we systematically study the frontier of reward maximization and KL minimization achieved by different alignment methods in a controlled text generation task following previous works (Ziegler et al., 2019; Rafailov et al., 2023). We investigate two different settings, including learning directly from preference data governed by a predefined oracle reward model, and 2) learning from a reward model estimated from the preference data. In §4.2, we compare EXO against different approaches on realistic alignment problems including generating human-preferred summaries, helpful dialogue responses, and answers that follow human instructions. Under both settings of learning directly from the preference data and from a reward model, EXO outperforms existing alignment approaches, e.g., DPO and PPO. Next, we briefly describe the experiment settings and leave additional details in Appendix B.

Experiment Setting. Our experiments consider two practical settings of LM alignment: (i) Directly train on a preference dataset $\mathcal{D}^{\text{pref}} = \{\boldsymbol{x}^{(j)}, \boldsymbol{y}_w^{(j)}, \boldsymbol{y}_l^{(j)}\}_{j=1}^N$ where $\boldsymbol{y}_w$ and $\boldsymbol{y}_l$ are the chosen and rejected responses judged by an oracle reward model or human labelers. (ii) Train on a reward dataset $\mathcal{D}^{\text{rw}} = \{\boldsymbol{x}^{(j)}, (\boldsymbol{y}_1^{(j)}, r_1^{(j)}), \cdots, (\boldsymbol{y}_K^{(j)}, r_K^{(j)})\}_{j=1}^N$ where $\boldsymbol{y}_k$ is generated by the SFT policy and r_k is a scalar reward provided by a reward model estimated on the given preference dataset. In the **controlled text generation** task, the policy is optimized to generate a completion $\boldsymbol{y}$ with positive sentiment given a prefix $\boldsymbol{x}$ of a movie review from the IMDB dataset[2] (Maas et al., 2011). To systematically evaluate the alignment performance, we train a binary sentiment classifier on the IMDB dataset and define the oracle reward as its log odds following Ziegler et al. (2019). Both the policy and the reward models are initialized from the GPT-2 large model (Radford et al., 2019). In the **summarization** task, the policy is required to generate a summary $\boldsymbol{y}$ of the post $\boldsymbol{x}$ from the Reddit forum that is preferred by human annotators. Following Stiennon et al. (2020), we use the same filtered version[3] of the Reddit TL;DR summarization dataset (Völske et al., 2017) to train the SFT policy and use their preference dataset[4] for the alignment problem. In the **dialogue generation** task, the policy is learned to generate a helpful response $\boldsymbol{y}$ given multi-turn dialogue history between the user and the assistant denoted as $\boldsymbol{x}$. We use the helpfulness subset of the Anthropic Helpful and Harmless dialogue dataset[5] (Bai et al., 2022) as the preference dataset and train the SFT policy using the chosen responses. For summarization and dialogue generation tasks, we initialize both the policy and the reward model from the Pythia-2.8B (Biderman et al., 2023) following Rafailov et al. (2023). To ensure sample quality, we use a temperature of $\tau = 0.8$ to divide the logits of the language model in all experiments. Lastly, for the **instruction following** task, we create a dataset based on instructions with high demand and representativeness from the real-world application scenarios, featuring 83K pairs of preferences annotated by human labelers and 49K prompts for policy training. The average lengths of the instructions and the answers are 47 and 230 respectively. We curate a diverse set of high-quality test instructions to assess a range of capabilities of the learned LM policy, encompassing multilingual ability, creative writing, open-ended question answering, and role playing. Each category takes the same proportion in the test set. Both the policy and the reward models are initialized from ChatGLM2-6B (Du et al., 2022).

Evaluation. In the controlled text generation task, we evaluate the frontier of the oracle reward and the KL divergence achieved by different approaches. This enables us to systematically compare the effectiveness of different methods in maxmizing the oracle reward under the same distributional shift constrained by the reverse KL. For experiments on the public preference datasets of summarization and dialogue generation, we use the reward model trained on the preference dataset as an in-domain proxy of the unknown ground-truth reward and also query GPT-4 for zero-shot pair-wise evaluation, which is shown to be consistent with human judgments (Rafailov et al., 2023). The prompts for GPT-4 evaluation are slightly modified based on those used in Rafailov et al. (2023), as detailed in Appendix B. We compare the generated outputs against those generated by the SFT policy and the preferred choice in the preference dataset. For the instruction-following task, we report the win rate of directly comparing our method against various baselines as judged by GPT-4. Additionally, we employ human assessment to evaluate criteria including adherence to instruction, correctness, fluency, safety and helpfulness.

Methods for Comparison. We consider the following methods for aligning language models with human prefer-

[2]`https://huggingface.co/datasets/imdb`.

[3]`https://huggingface.co/datasets/UCL-DARK/openai-tldr-filtered`.

[4]`https://huggingface.co/datasets/openai/summarize_from_feedback`.

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

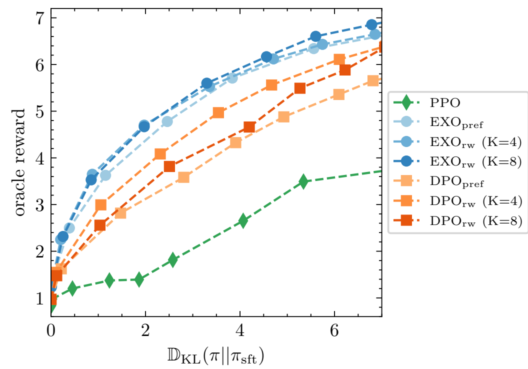


Figure 2. The frontier of oracle reward vs reverse KL to the SFT policy of different methods in the controlled experiment.

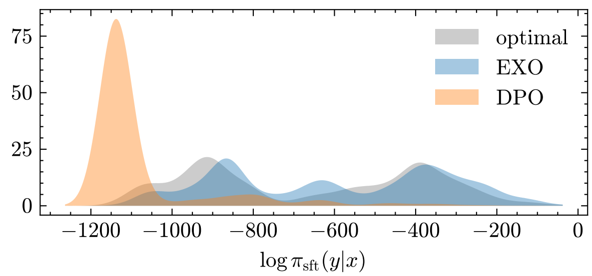


Figure 3. Visualization of the estimated density ratio between the optimal and learned policy by EXO and DPO and the SFT policy on samples from the SFT policy sorted by their log probabilities.

ences under various settings. Under the setting of learning directly from preferences, we consider the special case of EXO for preference learning $\mathcal{L}_{\text{exo-pref}}$ (Eq. (16)) denoted as **EXO$_{\text{pref}}$**, and the standard DPO (Rafailov et al., 2023) that minimizes the pair-wise loss $\mathcal{L}_{\text{dpo}}$ (Eq. (6)) on the preference data, which we denoted as **DPO$_{\text{pref}}$**. Then we consider the setup of alignment with a reward model estimated from the preference dataset, which includes the RL algorithm **PPO** (Ziegler et al., 2019) that optimizes the expected reward with a KL penalty (Eq. (3)), the general EXO objective $\mathcal{L}_{\text{exo}}$ (Eq. (13)) that performs probability matching by minimizing reverse KL, which is denoted as **EXO$_{\text{rw}}$**, the general DPO objective $\mathcal{L}_{\text{dpo-rw}}$ (Eq. (17)) that minimizes the forward KL, which is denoted as **DPO$_{\text{rw}}$**, and the **Best-of-N** method which first samples $N = 128$ outputs from the SFT policy and then returns the response with the highest score according to the reward model. Note that the Best-of-N baseline is practically inefficient and can be regarded as an upperbound of exploiting the SFT policy according to the reward model in Eq. (4) by trading off the computation.

4.1. Alignment with the Oracle Reward

To avoid undesirable reward overoptimizaion due to distributional shift, a preferred alignment solution should return a policy that obtains high oracle reward while incurring minimum deviation from π_{sft}. Thereby, we plot the frontier of the oracle reward against KL divergence in Figure 2. We additionally present the accuracy-KL frontier in Figure 7 in Appendix C.3. Each point represents a checkpoint of the learned policy which is evaluated on 512 prefixes from the test set to complete the response with maximumly 512 tokens. We sample $M = 4$ completions $\{\boldsymbol{y}_i\}_{i=1}^M$ for each given prompt $\boldsymbol{x}$ to calculate the average oracle reward as well as to reduce the variance of approximating the sequence-level KL divergence $\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{sft}}) \approx \frac{1}{M}\sum_{i=1}^M \log \pi_\theta(\boldsymbol{y}_i|\boldsymbol{x}) - \log \pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})$.

Despite aiming to optimize the same alignment objective, the EXO approaches (EXO$_{\text{pref}}$ and EXO$_{\text{rw}}$) yield the most efficient frontiers in their respective settings , evidenced by consistently achieving higher oracle rewards than baselines under the same KL divergence. Specifically, in the setting of directly learning from preference pairs, EXO$_{\text{pref}}$ outperforms DPO$_{\text{pref}}$ by a large margin, which underscores EXO's better sample efficiency when learning from a constrained source of preference indicators. As K increases, the frontier of EXO$_{\text{rw}}$ begins to exhibit convergence, indicating an effective progression towards the intended solution within a finite K. Although DPO$_{\text{rw}}$ also improves over DPO$_{\text{pref}}$ when $K = 4$, the frontier becomes worse when K is further increased to 8. This result substantiates our analysis about the mean-seeking behavior of forward KL, which leads to inaccuracy in capturing the modes of the complex target distribution. Finally, we illustrate the strong optimization efficiency of EXO in Figure 8, evidenced by consistently achieving high and stable oracle rewards within fewer number of training steps compared with PPO and DPO in Appendix C.4.

Next, we compare DPO and EXO from the probability matching perspective by visualizing the probability density of the policies obtained by these two approaches[6]. In Figure 3, we plot the estimated density ratio of the optimal and learned policies by EXO and DPO against π_{sft} given a randomly chosen test prompt "*This Fox spectacle was a big hit when released in*". Since the probability density of an LM policy is defined over a high dimensional space of $\mathcal{Y} = \mathcal{V}^n$, it is intractable to evaluate every point in this space exhaustively. Thus, we consider the representative data points that are sampled from π_{sft}, and sort them in the ascending order of their log probabilities. Then we compute the empirical distribution under the learned policies over these samples. Formally, given $M = 256$ samples $\{\boldsymbol{y}_i\}_{i=1}^M$ drawn from π_{sft} conditioned on the prompt $\boldsymbol{x}$, the empirical distribution $\hat{\pi}_\theta$ is calculated via self-normalized importance sampling over the learned policy π_θ:

$$\hat{\pi}_\theta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\sum_{j=1}^M \pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})/\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}.$$

[6]We consider the setting of learning directly from preferences.

Method	Reward Model (%)		GPT-4 (%)	
	vs SFT	vs Chosen	vs SFT	vs Chosen
`w/ Preferences`				
DPO_{pref}	68.3	23.7	57.0	30.5
EXO_{pref}	**92.5**	**60.1**	**83.0**	**55.0**
`w/ Reward Model`				
Best-of-N	99.3	75.8	83.5	60.0
PPO	93.2	58.3	77.0	52.0
DPO_{rw}	82.7	39.8	70.0	41.0
EXO_{rw}	**97.3**	**76.4**	**88.5**	**64.0**

Table 1. Win rates against the SFT generated texts and the chosen texts on the TL;DR summarization dataset. Best results from the computationally efficient methods are highlighted in **boldface**.

Method	Reward Model (%)		GPT-4 (%)	
	vs SFT	vs Chosen	vs SFT	vs Chosen
`w/ Preferences`				
DPO_{pref}	66.3	65.1	58.0	37.0
EXO_{pref}	**76.4**	**76.7**	**73.0**	**51.0**
`w/ Reward Model`				
Best-of-N	94.6	98.2	86.0	63.0
PPO	75.0	74.0	66.5	52.0
DPO_{rw}	79.9	81.3	75.5	49.0
EXO_{rw}	**85.6**	**87.2**	**83.5**	**60.0**

Table 2. Win rates against the SFT generated texts and the chosen texts on the Anthropic-HH dataset. Best results from the computationally efficient methods are highlighted in **boldface**.

For the optimal policy, the empirical distribution reduces to:

$$\hat{\pi}^*_\beta(\boldsymbol{y}_i|\boldsymbol{x}) = \frac{M\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})\exp(r(\boldsymbol{x},\boldsymbol{y}_i)/\beta)}{\sum_{j=1}^{M}\exp(r(\boldsymbol{x},\boldsymbol{y}_j)/\beta)}.$$

Finally, we use kernel density estimation to estimate the probability density $\hat{\pi}(\boldsymbol{y}|\boldsymbol{x})$ of the empirical distribution and plot the density ratio $\rho_{\hat{\pi}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\hat{\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$ against the log probability of the data points under π_{sft}. From the result, the density ratio of the EXO policy closely matches the optimal policy at the high probability region against π_{sft}, which reflects its ability in capturing the major modes of the optimal policy. However, the DPO policy overestimates the long tail in π_{sft} due to the mean-seeking tendency of forward KL, resulting in a reduced accuracy in fitting the modes of the optimal policy. We present more visualization results in Appendix C.5.

4.2. Alignment with Human Preferences

Next, we undertake a series of experiments focused on learning from real human preferences. In the tasks of summarization and dialogue generation, we consider the two settings of learning directly from preference data and from a reward model respectively. We set the same hyperparameters (e.g., β_π, β_r) for EXO and DPO across different settings and datasets, and provide the results of tuning these hyperparameters in Appendix C.1 and C.2. Evaluation results on text summaization and dialogue generation are shown in Table 1 and Table 2 respectively. Upon comparison with both the SFT and chosen responses, EXO demonstrates notable improvement over DPO and PPO. This advantage is evident in evaluations using both the in-domain reward model and zero-shot assessment by GPT-4, across both settings of learning from preferences and from a reward model. Notably, EXO is the only practically efficient method to attain a GPT-4 win rate exceeding 60% when compared to the chosen responses that may have been produced by a more advanced language model. Although the Best-of-N baseline achieves comparable or higher reward model win rate by maximizing out the reward model, it suffers from the most significant decline of win rate when assessed by GPT-4. This drop of performance is attributed to its excessive exploitation of the imperfect reward model while neglecting the deviation from the initial SFT policy. We provide examples generated by DPO and EXO on both tasks in Appendix C.6.

For the instruction-following task, we report the win rates of EXO compared to various baselines in Figure 4 under the setting of learning from the reward model given its advantageous performance observed so far. From the result, we observe that EXO outperforms all baselines by clear margins, thereby underscoring its scalability in practical applications. Notably, EXO achieves 10% and 5% improvement over its closest competitors as judged by GPT-4 and human annotators respectively.

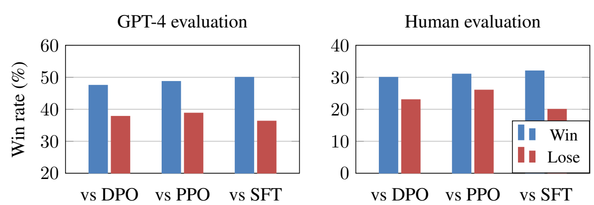


Figure 4. Win rates by comparing EXO to various baselines on the instruction-following task judged by GPT-4 and human labelers.

5. Related Work

Large language models (Rae et al., 2021; Hoffmann et al., 2022; Touvron et al., 2023) learned to predict the next tokens on large corpora have significantly improved the performance of various natural language processing (NLP) tasks in zero shot (Radford et al., 2019) or few-shot (Brown et al., 2020; Chowdhery et al., 2023) settings. To bridge the mismatch between the training objective and users' objective of solving specific NLP tasks, instruction-tuning is proposed, which fine-tunes the language models on human curated instruction-response pairs in a fully supervised manner (Sanh et al., 2021; Chung et al., 2022; Longpre et al., 2023; Wang et al., 2023). Despite its performance and generalization to unseen instruction sets (Chung et al., 2022), there have been concerns that the model learned to predict the next token might only capture surficial patterns rather than comprehending the tasks (Kung & Peng, 2023).

To address the aforementioned deficiencies, the framework of reinforcement learning from human feedback (RLHF) is proposed (Ziegler et al., 2019), which relies on only *relative* human preferences on response quality and optimizes the language model by RL algorithms (Williams, 1992), e.g., PPO (Schulman et al., 2017), under the supervision of a reward model which encapsulates the implicit preference of humans. Despite the success of RLHF in various tasks where alignment is strongly emphasized, e.g., translation (Kreutzer et al., 2018), summarization (Stiennon et al., 2020), instruction following (Ouyang et al., 2022), and etc., RL fine-tuning for large language models still faces serious challenges in stability and scalability (Zheng et al., 2023b).

Thereby, a new trend of recent works proposed to optimize the language model to follow human preferences without applying RL algorithms (Yuan et al., 2023; Zhao et al., 2023; Rafailov et al., 2023). While most approaches are empirically set to model alignment as a re-ranking problem, DPO proposed a theoretically sound way to realize direct policy optimization from preference data based on the analytic optimal solution of the reward maximization objective in RLHF. Afterwards, there have been several studies that extend DPO in various ways, e.g., altering the preference data by sampling from the optimal policy via rejection sampling (Liu et al., 2023), substituting the point-wise reward with a pair-wise preference function (Azar et al., 2023), extending the preference pairs to rankings of preferences of any size (Song et al., 2023), and etc. However, these approaches are still based on the formation of DPO. In this work, we propose to learn the policy by exactly optimizing the RLHF objective via probability matching that minimizes the reverse KL. In the literature of maximum-entropy RL (Eysenbach & Levine, 2019; Korbak et al., 2022), this equivalent form of probability matching was discussed only for analysis purposes, while we are the first to derive a practical way to optimize it. We also revisit DPO under this perspective and recognize that it actually corresponds to minimizing the forward KL in its general form. To this sense, DPO shares the same spirit of weighted regression (Peters & Schaal, 2007; Peng et al., 2019; Wang et al., 2020), an algorithm that directly utilizes behavioral actions to supervise the policy in offline RL (Peters et al., 2010; Lu et al., 2023). However, this approach is known to be suboptimal when the policy model is limited in distributional expressivity (Yue et al., 2022; Chen et al., 2023). We analyze the characteristics of the probability density learned by DPO with both theoretical insight and empirical experimentations.

6. Limitations and Future Work

The alignment framework proposed by Ziegler et al. (2019) relies on the KL regularization to the SFT policy to prevent the optimized policy from greedily maximizing out the reward model which is estimated from the human preference data. Despite the regularization, our experiments still reveal instances of reward over-optimization, a phenomenon possibly due to insufficient focus on the reward model estimation and use. Importantly, ensuring that the reward model accurately reflects the true modes of the oracle distribution of human is more vital for PPO and EXO that optimizes a mode-seeking objective. This opens possible avenues to improve the alignment framework at a broader scope. For instance, rather than relying on a static preference dataset for reward model training, it could be more effective to dynamically improve the reward model with the development of the policy, thereby offering more precise feedback. Upon the current regularization that solely focuses on proximity to the initial policy, one can take into account the uncertainty of the reward model output to avoid over-exploitation of the reward model. It is beneficial to take into account these aspects to develop efficient and effective method towards closer alignment with human preferences. Additionally, while we already evaluated EXO on advanced language models up to 6B on realistic scenarios, scaling EXO to models that are orders of magnitude larger can present profound implications. At the other end of the spectrum, systematically dissecting and comparing PPO, DPO and EXO, e.g., regarding their variance and bias during optimization is essential to broaden our understanding of these methods.

7. Conclusion

In this work, we consider the problem of aligning language models with human preferences. Although reinforcement learning (RL) for reward maximization presents the direct and apparent solution, we reframe the problem in a supervised probability matching framework, which underscores the probabilistic interpretation of the alignment procedure. This derives our *efficient exact optimization* (EXO) of the KL-regularized reward maximization objective of alignment. Formally, we prove the asymptotic equivalence between the EXO objective and the alignment objective. In practice, EXO enables efficient optimization via probability matching between empirical distributions, which avoids the complexities of RL algorithms. We further demonstrate that DPO in its general form actually corresponds to minimizing the forward KL against the optimal policy, which is shown to be less effective in capturing the modes of the optimal policy under realistic model parametrization with both theoretical and empirical justifications. Finally, we demonstrate the effectiveness and scalability of EXO on various text generation tasks with real human preferences.

Acknowledgements

This work was supported by the NSFC projects (with No. 61936010 and No. 62306160). This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604). This work was also supported by China National Postdoctoral Program for Innovative Talents (No. BX20230194) and China Postdoctoral Science Foundation (No. 2023M731952). We would also like to thank Zhipu AI for sponsoring the computation resources and annotation cost used in this work.

Impact Statement

This paper presents a method whose goal is to advance the alignment of language models with human preferences. This endeavor, while technically challenging, carries significant implications for the ethical use and societal impact of artificial intelligence. The goal of alignment aims to mitigate the inherent biases of AI systems and ensure that they reflect diverse human values, goals and intentions that are safe and ethical. It enhances AI's utility in various sectors to make reliable decisions that are in line with organizational goals and ethical standards.

References

Anschel, O., Baram, N., and Shimkin, N. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International conference on machine learning*, pp. 176–185. PMLR, 2017.

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences. *CoRR*, abs/2310.12036, 2023. doi: 10.48550/ARXIV.2310.12036. URL https://doi.org/10.48550/arXiv.2310.12036.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., Showk, S. E., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Chan, A., Silva, H., Lim, S., Kozuno, T., Mahmood, A. R., and White, M. Greedification operators for policy optimization: Investigating forward and reverse kl divergences. *The Journal of Machine Learning Research*, 23 (1):11474–11552, 2022.

Chen, H., Lu, C., Ying, C., Su, H., and Zhu, J. Offline reinforcement learning via high-fidelity generative behavior modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=42zs3qa2kpy.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. GLM: general language model pretraining with autoregressive blank infilling. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 320–335. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.26. URL https://doi.org/10.18653/v1/2022.acl-long.26.

Eysenbach, B. and Levine, S. If maxent rl is the answer, what is the question? *arXiv preprint arXiv:1910.01913*, 2019.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Ji, H., Ke, P., Hu, Z., Zhang, R., and Huang, M. Tailoring language generation models under total variation distance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/pdf?id=VELL0PlWfc.

Ji, H., Ke, P., Wang, H., and Huang, M. Language model decoding as direct metrics optimization. *arXiv preprint arXiv:2310.01041*, 2023b.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023c.

Korbak, T., Perez, E., and Buckley, C. L. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022.

Kreutzer, J., Uyheng, J., and Riezler, S. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. *arXiv preprint arXiv:1805.10627*, 2018.

Kung, P.-N. and Peng, N. Do models really learn to follow instructions? an empirical study of instruction tuning. *arXiv preprint arXiv:2305.11383*, 2023.

Lin, C.-C., Jaech, A., Li, X., Gormley, M., and Eisner, J. Limitations of autoregressive models and their alternatives. In *NAACL*, 2021.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Lu, C., Chen, H., Chen, J., Su, H., Li, C., and Zhu, J. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22825–22855. PMLR, 2023. URL https://proceedings.mlr.press/v202/lu23d.html.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R. (eds.), *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 142–150. The Association for Computer Linguistics, 2011. URL https://aclanthology.org/P11-1015/.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019. URL http://arxiv.org/abs/1910.00177.

Peters, J. and Schaal, S. Reinforcement learning by reward-weighted regression for operational space control. In Ghahramani, Z. (ed.), *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pp. 745–750. ACM, 2007. doi: 10.1145/1273496.1273590. URL https://doi.org/10.1145/1273496.1273590.

Peters, J., Mülling, K., and Altun, Y. Relative entropy policy search. In Fox, M. and Poole, D. (eds.), *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*, pp. 1607–1612. AAAI Press, 2010. doi: 10.1609/AAAI.V24I1.7727. URL https://doi.org/10.1609/aaai.v24i1.7727.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, H. F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., van den Driessche, G., Hendricks, L. A., Rauh, M., Huang, P., Glaese, A., Welbl, J., Dathathri, S., Huang, S., Uesato, J., Mellor, J., Higgins, I., Creswell, A., McAleese, N., Wu, A., Elsen, E., Jayakumar, S. M., Buchatskaya, E., Budden, D., Sutherland, E., Simonyan, K., Paganini, M., Sifre, L., Martens, L., Li, X. L., Kuncoro, A., Nematzadeh, A., Gribovskaya, E., Donato, D., Lazaridou, A., Mensch, A., Lespiau, J., Tsimpoukelli, M., Grigorev, N., Fritz, D., Sottiaux, T., Pajarskas, M., Pohlen, T., Gong, Z., Toyama, D., de Masson d'Autume, C., Li, Y., Terzi, T., Mikulik, V., Babuschkin, I., Clark, A., de Las Casas, D., Guy, A., Jones, C., Bradbury, J., Johnson, M. J., Hechtman, B. A., Weidinger, L., Gabriel, I., Isaac, W., Lockhart, E., Osindero, S., Rimell, L., Dyer, C., Vinyals, O., Ayoub, K., Stanway, J., Bennett, L., Hassabis, D., Kavukcuoglu, K., and Irving, G. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

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

Song, F., Bowen, Y., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. *ArXiv*, abs/2306.17492, 2023. URL https://api.semanticscholar.org/CorpusID:259308873.

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

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl;dr: Mining reddit to learn automatic summarization. In Wang, L., Cheung, J. C. K., Carenini, G., and Liu, F. (eds.), *Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pp. 59–63. Association for Computational Linguistics, 2017. doi: 10.18653/V1/W17-4508. URL https://doi.org/10.18653/v1/w17-4508.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.754. URL https://doi.org/10.18653/v1/2023.acl-long.754.

Wang, Z., Novikov, A., Zolna, K., Merel, J., Springenberg, J. T., Reed, S. E., Shahriari, B., Siegel, N. Y., Gülçehre, Ç., Heess, N., and de Freitas, N. Critic regularized regression. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/588cb956d6bbe67078f29f8de420a13d-Abstract.html.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Yao, Z., Aminabadi, R. Y., Ruwase, O., Rajbhandari, S., Wu, X., Awan, A. A., Rasley, J., Zhang, M., Li, C., Holmes, C., Zhou, Z., Wyatt, M., Smith, M., Kurilenko, L., Qin, H., Tanaka, M., Che, S., Song, S. L., and He, Y. Deepspeed-chat: Easy, fast and affordable RLHF training of chatgpt-like models at all scales. *CoRR*, abs/2308.01320, 2023. doi: 10.48550/ARXIV.2308.01320. URL https://doi.org/10.48550/arXiv.2308.01320.

Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Yue, Y., Kang, B., Ma, X., Xu, Z., Huang, G., and YAN, S. Boosting offline reinforcement learning via data rebalancing. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022. URL https://openreview.net/forum?id=vOC01fqW2T.

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685, 2023a. doi: 10.48550/ARXIV.2306.05685. URL https://doi.org/10.48550/arXiv.2306.05685.

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P. F., and Irving, G. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593, 2019. URL http://arxiv.org/abs/1909.08593.

A. Proofs and Derivations

A.1. Proof of Theorem 3.1

Proof. We derive the optimal π_{θ^*} that maximizes the generalized alignment objective $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ and show that it equals to the optimal solution π_β^* of $\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta)$ given unlimited model capacity. First, we restate the formation of $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$:

$$\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r \mathbb{D}_{\text{KL}}[\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \| \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big), \tag{19}$$

where $\pi_\theta^{\beta_\pi}$ is defined as:

$$\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}. \tag{20}$$

Then we substitute θ with the optimal θ^* in Eq. (20) where π_{θ^*} maximizes $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$:

$$\pi_{\theta^*}^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_{\theta^*}(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}. \tag{21}$$

Since $\pi_{\theta^*}^{\beta_\pi}$ is also the optimal policy that maximizes $\mathcal{J}_{\text{lhf}}^{\beta_r}(\cdot)$, it should satisfy Eq. (4) which gives:

$$\pi_{\theta^*}^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) = \pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r(\boldsymbol{x},\boldsymbol{y})}. \tag{22}$$

Together with Eq. (21) and Eq. (22), we obtain the form of π_{θ^*} via some simple algebra:

$$\begin{aligned}
\pi_{\theta^*}(\boldsymbol{y}|\boldsymbol{x}) &\propto (\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi-1})^{\frac{1}{\beta_\pi}} \\
&\propto (\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r(\boldsymbol{x},\boldsymbol{y})})^{\frac{1}{\beta_\pi}}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{\frac{\beta_\pi-1}{\beta_\pi}} \\
&\propto \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{\frac{1}{\beta_\pi}}e^{\frac{1}{\beta_r\beta_\pi}r(\boldsymbol{x},\boldsymbol{y})}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{\frac{\beta_\pi-1}{\beta_\pi}} \\
&\propto \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r\beta_\pi}r(\boldsymbol{x},\boldsymbol{y})}.
\end{aligned}$$

By definition, $\beta_r\beta_\pi = \beta$, then π_{θ^*} reduces to the same form of the optimal solution of $\mathcal{J}_{\text{lhf}}^{\beta}(\pi_\theta)$ defined in Eq. (4):

$$\pi_{\theta^*}(\boldsymbol{y}|\boldsymbol{x}) = \pi_\beta^*(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta}r(\boldsymbol{x},\boldsymbol{y})},$$

which completes the proof. □

A.2. Derivation of Eq. (9)

We first start by rearranging $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$ into the expectation of a log ratio:

$$\begin{aligned}
\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r\mathbb{D}_{\text{KL}}[\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\|\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})]\Big) \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}[r_\phi(\boldsymbol{x},\boldsymbol{y})] - \beta_r\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big[\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\Big]\Big) \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\Big(\beta_r\mathbb{E}_{\boldsymbol{y}\sim\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big[\log e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}\Big] - \beta_r\mathbb{E}_{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big[\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\Big]\Big) \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\boldsymbol{y}\sim\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big[\beta_r\log\frac{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}}{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big]
\end{aligned}$$

Notice the analytical form of $\pi_{\beta_r}^*$:

$$\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\beta_r(\boldsymbol{x})}}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}.$$

We substitute $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}$ into the expression of $\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi})$:

$$\begin{aligned}
\mathcal{J}_{\text{lhf}}^{\beta_r}(\pi_\theta^{\beta_\pi}) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\boldsymbol{y}\sim\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big[\beta_r\log\frac{Z_{\beta_r(\boldsymbol{x})}\pi_{\beta_r}^*(\boldsymbol{y}|\boldsymbol{x})}{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}\Big] \\
&= \beta_r\mathbb{E}_{\boldsymbol{x}\in\mathcal{D}^{\text{pref}}}\Big[-\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}(\cdot|\boldsymbol{x})\|\pi_{\beta_r}^*(\cdot|\boldsymbol{x})) + \log Z_{\beta_r}(\boldsymbol{x})\Big].
\end{aligned}$$

A.3. Proof of Theorem 3.2

Proof. We first restate the definition of $\mathcal{L}_{\text{exo}}(\pi_\theta)$ by substituting Eq. (10), (12) into (13):

$$\mathcal{L}_{\text{exo}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\sum_{i=1}^{K}\frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}}\Big(\log\frac{e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{f_\theta(\boldsymbol{x},\boldsymbol{y}_j)}} - \log\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\Big)\right] \tag{23}$$

Since $f_\theta(\boldsymbol{x},\boldsymbol{y}) = \log\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) - \log\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$, we have that:

$$\begin{aligned}\mathcal{L}_{\text{exo}}(\pi_\theta) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\sum_{i=1}^{K}\frac{e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^{K}e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}}\Big(\log\frac{e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^{K}e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}} - \log\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\Big)\right]\\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\sum_{i=1}^{K}\frac{\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}{\sum_{j=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}\Big(\log\frac{\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}{\sum_{j=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}} - \log\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\Big)\right].\end{aligned} \tag{24}$$

Since $\{\boldsymbol{y}_i\}_{i=1}^K$ are sampled from $\pi_{\text{sft}}(\cdot|\boldsymbol{x})$, when $K\to\infty$, for arbitary function $g:\mathcal{Y}\to\mathbb{R}$, the estimate $\frac{1}{K}\sum_{i=1}^K g(\boldsymbol{y}_i)$ is unbiased, i.e., $\lim_{K\to\infty}\frac{1}{K}\sum_{i=1}^K g(\boldsymbol{y}_i) = \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}[g(\boldsymbol{y})]$. We consider the following two instantiations of $g(\cdot)$.

For $g(\boldsymbol{y}) = \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}$, we have:

$$\begin{aligned}\sum_{j=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})} &= K\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\right]\\ &= K\sum_{\boldsymbol{y}\in\mathcal{Y}}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\\ &= K\sum_{\boldsymbol{y}\in\mathcal{Y}}\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\\ &= K.\end{aligned}$$

For $g(\boldsymbol{y}) = e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}$, we have:

$$\begin{aligned}\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)} &= K\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\left[e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}\right]\\ &= K\sum_{\boldsymbol{y}\in\mathcal{Y}}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}\\ &= KZ_{\beta_r}(\boldsymbol{x}).\end{aligned}$$

Then we simplify $\mathcal{L}_{\text{exo}}$ by substituting the expression of $\sum_{j=1}^K\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}$ and $\sum_{j=1}^K e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}$ when $K\to\infty$ into Eq. (24).

$$\begin{aligned}\mathcal{L}_{\text{exo}}(\pi_\theta) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\sum_{i=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{K\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}\Big(\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{K\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})} - \log\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{KZ_{\beta_r}(\boldsymbol{x})}\Big)\right]\\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}/Z_{\beta_r}(\boldsymbol{x})}\right].\end{aligned}$$

Notice the analytic form of $\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\beta_r}(\boldsymbol{x})}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y})}$, we substitute $\pi^*_{\beta_r}$ into the above equation:

$$
\begin{aligned}
\mathcal{L}_{\text{exo}}(\pi_\theta) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}\left[\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\frac{1}{K}\sum_{i=1}^{K}\sum_{\boldsymbol{y}_i\in\mathcal{Y}}\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\left[\sum_{\boldsymbol{y}\in\mathcal{Y}}\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x})}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}[\mathbb{D}_{\text{KL}}(\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})\|\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}))].
\end{aligned}
$$

Accordingly, by taking the derivative with respect to θ, we complete the proof of Eq. (14).

To prove Eq. (15), we utilize Eq. (14) to substitute into Eq. (9) to build the connection between $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi_\theta^{\beta_\pi})$ and $\mathcal{L}_{\text{exo}}$:

$$
\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi_\theta^{\beta_\pi}) = -\beta_r\mathcal{L}_{\text{exo}}(\pi_\theta) + \beta_r\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}[\log Z_{\beta_r}(\boldsymbol{x})].
$$

Then we take the gradient with respect to the parameters θ of the above formulat:

$$
\nabla_\theta\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi_\theta^{\beta_\pi}) = -\beta_r\nabla_\theta\mathcal{L}_{\text{exo}}(\pi_\theta),
$$

which completes the proof of Eq. (15). □

A.4. Proof of Theorem 3.3

Proof. We utilize the definition of $\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x}) \propto \pi_\theta(\boldsymbol{y}|\boldsymbol{x})^{\beta_\pi}\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})^{1-\beta_\pi}$ in Eq. (8) and divide both sides by $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})$:

$$
\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} \propto \left(\frac{\pi_\theta(\boldsymbol{y}|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})}\right)^{\beta_\pi}
$$

Then we substitute the above equation into $\mathcal{L}_{\text{dpo-rw}}$:

$$
\begin{aligned}
\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[-\sum_{i=1}^{K}\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\log\frac{e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^{K}e^{\beta_\pi\log\frac{\pi_\theta(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[-\sum_{i=1}^{K}\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\log\frac{e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}}{\sum_{j=1}^{K}e^{\log\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}}\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}}\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})}\left[-\sum_{i=1}^{K}\frac{e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}}\log\frac{\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})}}{\sum_{j=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}}\right]
\end{aligned}
\tag{25}
$$

Since $\{\boldsymbol{y}_i\}_{i=1}^K$ are sampled from $\pi_{\text{sft}}(\cdot|\boldsymbol{x})$, when $K\to\infty$, we follow the proof of Theorem 3.2 and directly give the following results:

$$
\begin{aligned}
&\sum_{j=1}^{K}\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})} = K, \\
&\sum_{j=1}^{K}e^{\frac{1}{\beta_r}r_\phi(\boldsymbol{x},\boldsymbol{y}_j)} = KZ_{\beta_r}(\boldsymbol{x}).
\end{aligned}
$$

Then we simplify $\mathcal{L}_{\text{dpo-rw}}$ by substituting the above results of $\sum_{j=1}^{K} \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_j|\boldsymbol{x})}{\pi_{\text{sft}}(\boldsymbol{y}_j|\boldsymbol{x})}$ and $\sum_{j=1}^{K} e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_j)}$ when $K \to \infty$ into Eq. (25):

$$\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})} \left[-\sum_{i=1}^{K} \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{K \pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})} \right]$$

Notice the analytic form of $\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\beta_r}(\boldsymbol{x})} \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x}) e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y})}$, we rearrange $\pi^*_{\beta_r}$ and substitute $\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x}) = \pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}) Z_{\beta_r}(\boldsymbol{x}) e^{-\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y})}$ into the above equation to simplify it:

$$\begin{aligned}
\mathcal{L}_{\text{dpo-rw}}(\pi_\theta) &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})} \left[-\sum_{i=1}^{K} \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \log \left(\frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})} \cdot \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \right) \right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})} \left[-\sum_{i=1}^{K} \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})} - \sum_{i=1}^{K} \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \log \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{K Z_{\beta_r}(\boldsymbol{x})} \right]
\end{aligned}$$

The second term of the final equality remains constant in relation to θ, and thus can be omitted when computing the derivative with respect to θ. Then we further consider the gradient of $\mathcal{L}_{\text{dpo-rw}}$:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\text{dpo-rw}}(\pi_\theta) &= \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \left[-\frac{1}{K} \sum_{i=1}^{K} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_i|\boldsymbol{x})} \left[\frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y}_i)}}{Z_{\beta_r}(\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}_i|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}_i|\boldsymbol{x})} \right] \right] \\
&= \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \left[-\mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x})} \left[\frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y})}}{Z_{\beta_r}(\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x})} \right] \right] \\
&= \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \left[-\sum_{\boldsymbol{y}\in\mathcal{Y}} \pi_{\text{sft}}(\boldsymbol{y}|\boldsymbol{x}) \frac{e^{\frac{1}{\beta_r} r_\phi(\boldsymbol{x},\boldsymbol{y})}}{Z_{\beta_r}(\boldsymbol{x})} \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x})} \right] \\
&= \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \left[-\sum_{\boldsymbol{y}\in\mathcal{Y}} \pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{\pi_\theta^{\beta_\pi}(\boldsymbol{y}|\boldsymbol{x})}{\pi^*_{\beta_r}(\boldsymbol{y}|\boldsymbol{x})} \right] \\
&= \nabla_\theta \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \left[\mathbb{D}_{\text{KL}}(\pi^*_{\beta_r}(\cdot|\boldsymbol{x}) \| \pi_\theta^{\beta_\pi}(\cdot|\boldsymbol{x})) \right],
\end{aligned}$$

which completes the proof of Theorem 3.3. □

A.5. Mechanistic Understanding of $\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta)$

We present the gradient of $\mathcal{L}_{\text{exo}}(\pi_\theta)$ defined in Eq. (13):

$$\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta) = -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}^{\text{pref}}} \mathbb{E}_{\pi_{\text{sft}}(\boldsymbol{y}_{1:K}|\boldsymbol{x})} \left[\sum_{k=1}^{K} p_{f_\theta}(k) \Big(\underbrace{\mathbb{D}_{\text{KL}}(p_{f_\theta} \| p_{r_\phi})}_{\textit{distributional}\text{ gap}} - \underbrace{\log \frac{p_{f_\theta}(k)}{p_{r_\phi}(k)}}_{\textit{point-wise}\text{ correction}} \Big) \nabla_\theta \log \pi_\theta(\boldsymbol{y}_k|\boldsymbol{x}) \right], \tag{26}$$

where $p_{f_\theta}(k)$ and $p_{r_\phi}(k)$ are short for $p_{f_\theta}(k|\boldsymbol{y}_{1:K}, \boldsymbol{x})$ and $p_{r_\phi}(k|\boldsymbol{y}_{1:K}, \boldsymbol{x})$, which are defined in Eq. (11) and Eq. (12) respectively.

Next, we provide a mechanistic understanding of $\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta)$, which is the expected weighted sum of the gradients of the log likelihood on samples $\boldsymbol{y}_{1:K}$ drawn from $\pi_{\text{sft}}(\cdot|\boldsymbol{x})$. The weight is proportional to the difference between the log probability ratio $\log \frac{p_{f_\theta}(k)}{p_{r_\phi}(k)}$ and the KL divergence $D_{\text{KL}}(p_{f_\theta} \| p_{r_\phi})$. Intuitively, if the policy has already correctly weighted the sample y_k according to the reward model, i.e., $p_{r_\phi}(k) = p_{f_\theta}(k)$, then $\log \frac{p_{f_\theta}(k)}{p_{r_\phi}(k)} = 0$ and the weight suggests that it only needs to minimize the overall KL divergence between p_{f_θ} and p_{r_ϕ} on the distribution level. If the policy π_θ overestimates or underestimates the sample y_k, i.e., $\log \frac{p_{f_\theta}(k)}{p_{r_\phi}(k)} > 0$ or $\log \frac{p_{f_\theta}(k)}{p_{r_\phi}(k)} < 0$, this log-ratio will be used to calibrate the KL divergence to penalize or encourage the policy to update towards increasing the likelihood of this sample at a faster rate.

Finally, this gradient form offers us the insight of when the optimization should stop: the gradient $\nabla_\theta \mathcal{L}_{\text{exo}}(\pi_\theta)$ becomes 0 when the two distribution p_{f_θ} and p_{r_ϕ} are identical.

A.6. Illustrating the Relationship among the Objectives in §3

We illustrate the relationship among the objectives $\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta)$, $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$, $\mathcal{L}_{\text{dpo}}(\pi_\theta)$, $\mathcal{L}_{\text{dpo-rw}}(\pi_\theta)$ and $\mathcal{L}_{\text{exo}}(\pi_\theta)$ in Figure 5.

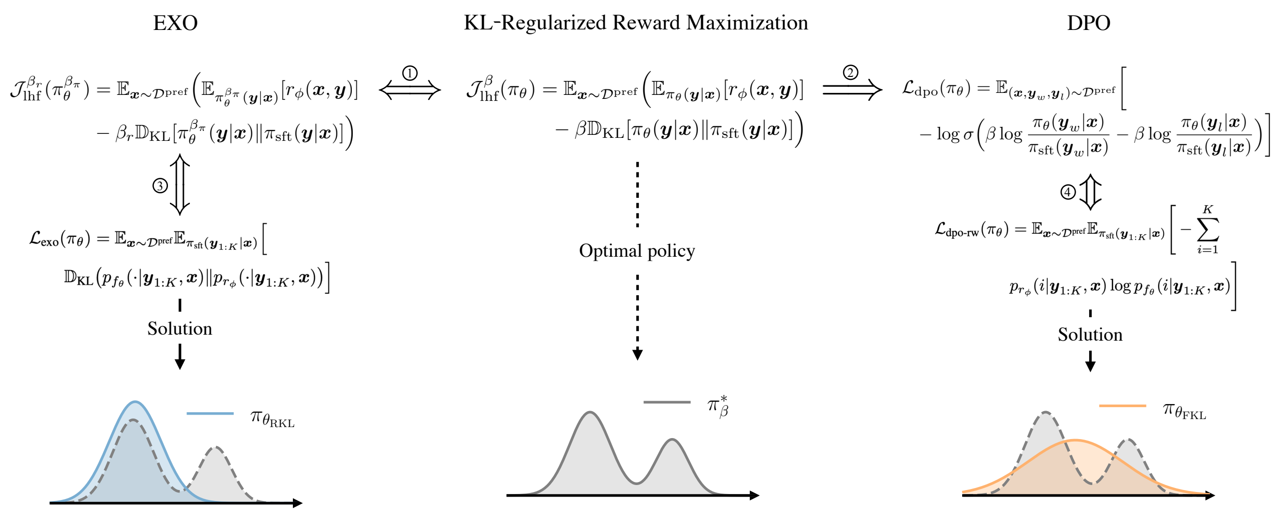


Figure 5. Illustration of the relationship among the different objectives discussed in §3. ①: $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$ is a generalized version of $\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta)$ by distributing the KL regularization to both the learned policy π_θ and the reward model r_ϕ (§3.1). ②: $\mathcal{L}_{\text{dpo}}(\pi_\theta)$ is derived based on the optimal policy of $\mathcal{J}^{\beta}_{\text{lhf}}(\pi_\theta)$ (§2.3). ③: $\mathcal{L}_{\text{exo}}(\pi_\theta)$ is equivalent to $\mathcal{J}^{\beta_r}_{\text{lhf}}(\pi^{\beta_\pi}_\theta)$ in terms of their optimization directions (§3.2). ④: $\mathcal{L}_{\text{dpo-rw}}$ is the generalized version of $\mathcal{L}_{\text{dpo}}$ by subsituting the pariwise loss with softmax loss over K responses.(§3.3). The optimal policy, denoted by a dotted line, assumes unlimited model capacity. The solution, shown with a solid line, is the practically achievable policy within the realistic constraints of model capacity.

B. Training and Evaluation Details

Training. In the controlled text generation task, we use $\beta_\pi = 0.1$ for EXO$_{\text{pref}}$ and DPO$_{\text{pref}}$, and additionally use $\beta_r = 0.1$ and $K \in \{4, 8\}$ for EXO$_{\text{rw}}$ and DPO$_{\text{rw}}$. For the tasks of summarization, dialogue generation, and instruction following, we use $\beta_\pi = 0.5$ for EXO$_{\text{pref}}$ and DPO$_{\text{pref}}$, and additionally use $\beta_r = 0.1$ and $K = 4$ for EXO$_{\text{rw}}$ and DPO$_{\text{rw}}$. We provide additional results of tuning β_r and β_π in Appendix C.1 to justify our choice of hyperparameters. In all experiments, we set the label smoothing hyperparameter ε in EXO$_{\text{pref}}$ to `1e-3`. For DPO and EXO, we use the Adam optimizer with a universal learning rate of `1e-6` and a batch size of 64 and train for one epoch on each dataset, although both methods converge within one epoch. All the hyperparameters are set to be the same for DPO and EXO for a fair comparison. For PPO, we modify based on the implementation of DeepSpeed-Chat (Yao et al., 2023), which sets $\gamma = 1$ and $\beta = 0.1$ by default. We pretrain the critic model for the first few steps while freezing the actor model and find it to improve convergence. We empirically tune the number of actor-freezing steps, total training steps, learning rate for actor and critic model, and the batch size for PPO on each dataset, as PPO is sensitive to these hyperparameters. Specifically, we conduct 15 trials of hyperparameter search on the IMDB dataset and 10 trials on the real human preference datasets in total. In the experiment, we report the PPO performance with the best hyperparameters obtained under constrained number of hyperparameter search trials. We conduct the experiments except for instruction following on 8 V100 GPUs. For instruction following task, we train the models on 8 A100 GPUs.

Evaluation. At inference time, we sample 4 completions from the learned policy for each prompt and consider 512 prompts from the test set for all datasets. Except for the instruction following task, we sample from the policy with the same temperature $\tau = 0.8$ that is set during training for consistency. For the instruction following task, we use top-p sampling and empirically set $p = 0.9$ and temperature $\tau = 0.95$ given its performance. To calculate the win rate evaluated by the reward model, we consider all combinations of pairs between the completions generated by the learned policy and the base completions (either generated by the SFT policy or the chosen completion in the dataset) and then compare the scores from the reward model on the pairs of generations. For the evaluations using GPT-4, we sample 100 prompts and 1 completion for each prompt under each policy. To mitigate the position bias of GPT-4, we evaluate one pair of generations twice by swapping the order of responses in each pair. To evaluate the quality of the summaries, we use the concise prompt of Rafailov et al. (2023) as shown in Table 3. To evaluate the helpfulness of the generated dialogues, we use the prompt shown in Table 4, which is modified based on the prompt of Rafailov et al. (2023) for single-turn dialogue to accommodate the general multi-turn setting. For the instruction-following task, we use the prompt modified from the prompt for reference-guided pairwise comparison provided in Zheng et al. (2023a).

Human Assessment. We conduct human assessment to evaluate the instruction following task more thoroughly. Specifically, we select three matchups that pair the generated outputs of EXO with those produced by DPO, PPO and the SFT policy. Given 100 randomly sampled test instructions, each model generates 100 responses, which results in a total of 300 pairs of comparisons. We assign 3 human labelers to each comparison, producing 900 judgements in total. Given the instruction, each human annotator is provided with two generated answers by two systems respectively together with a high-quality reference answer. The annotator is then asked to make a preference among win, tie or lose by comparing the generated answers with the reference answer, considering the criteria including adherence to instruction, correctness, fluency, safety and helpfulness. Specifically, adherence to instruction encapsulates the model's comprehension and following of the prompt's intention. Correctness involves the identification of inaccurate knowledge or logical inconsistencies within the generated responses. Fluency assesses the linguistic coherence, encompassing an examination of sentence completeness, grammatical accuracy, and the presence of a consistent language structure. Safety refers to the inspection for potentially harmful content. Lastly, helpfulness indicates whether the responses provide the information required by the prompt or contribute to problem resolution.

```
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 A:
<Summary A>

Summary B:
<Summary B>

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

Table 3. Prompt for GPT-4 evaluation on the summarization task. Texts in blue are placeholders to be substituted by the real data.

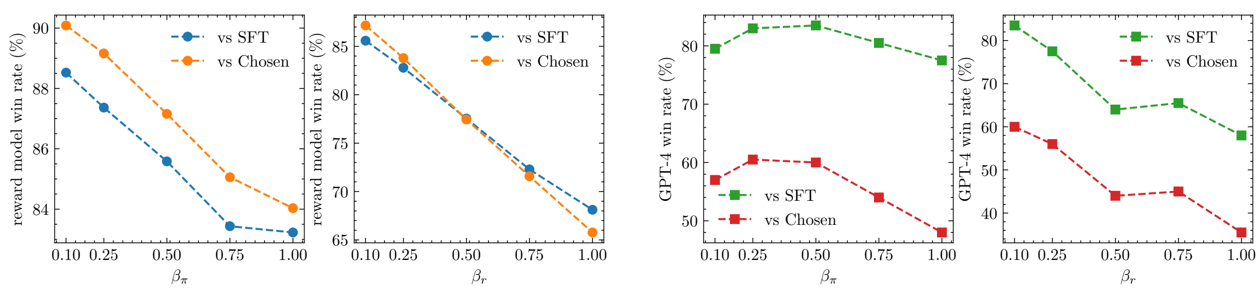
C. Additional Experiment Results

C.1. Ablation Study of β_r and β_π

We present an ablation study to investigate the performance of EXO_{rw} on the dialogue generation task by varying β_r and β_π respectively. We execute multiple runs bifurcated into two series. We set $\beta_\pi = 0.5$ as the default value and vary $\beta_r \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$. Subsequently, the process is reversed whereby we fix $\beta_r = 0.1$, and alter $\beta_\pi \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$. We present the results in Figure 6.

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

Dialogue history:
<dialogue history>

Response A:
<Response A>

Response B: <Response B>

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

Table 4. Prompt for GPT-4 evaluation on the dialogue generation task. Texts in blue are placeholders to be substituted by the real data.



Reward model win rate. GPT-4 win rate.

Figure 6. Win rate performance of EXO_{rw} evaluated by the reward model and the GPT-4 by varying β_r and β_π.

From Figure 6 (a), we observe that reducing either β_r and β_π increase the reward model win rate, which is expected as a small $\beta_r \times \beta_\pi$ encourages the policy to optimize the reward model while neglecting the KL regularization. In Figure 6 (b), the GPT-4 win rate starts to decline when $\beta_\pi < 0.5$, which suggests where the reward model starts to be over-optimized. The different effects of tuning β_π and β_r on the performance could be attributed to their different roles in our algorithm, i.e., β_π scales the log probability in the parametrized policy while β_r scales the reward. Based on the results, we recommend adopting a moderate value for β_π and a lower value for β_r, for instance, $\beta_\pi = 0.5$ and $\beta_r = 0.1$.

C.2. Effect of β_r and β_π beyond the Product $\beta = \beta_r\beta_\pi$

To further demonstrate the effect of β_r and β_π beyond the impact on their product $\beta = \beta_\pi\beta_r$, we fix $\beta = 0.05$ while tuning $\beta_\pi \in \{1, 0.5, 0.1, 0.05\}$ and $\beta_r \in \{0.05, 0.1, 0.5, 1\}$ accordingly. The result is shown in Table 5. According to the evaluation by GPT-4, the best performnce is obtained at $\beta_\pi = 0.5, \beta_r = 0.1$, while neither the hyper-parameter choice of $\beta_\pi = 1, \beta_r = 0.05$ (the configuration of PPO) nor the choice of $\beta_\pi = 0.05, \beta_r = 1$ (the configuration of DPO) yield the highest win rate evaluated by GPT-4. While using a small β_π is more likely to trigger over-optimization of the reward model, which indicates the asymmetric effect of the two hyperparameters β_r and β_π on the optimization process.

β_r	β_π	Reward Model (%) vs SFT	Reward Model (%) vs Chosen	GPT-4 (%) vs SFT	GPT-4 (%) vs Chosen
1.0	0.05	87.1	88.4	81.0	57.5
0.5	0.1	87.1	88.0	77.0	59.5
0.1	0.5	85.6	87.2	**83.5**	**60.0**
0.05	1.0	84.1	84.8	76.5	52.0

Table 5. Effect of β_r and β_π beyond β by tuning β_π and β_r while keeping their product $\beta = \beta_r\beta_\pi$ fixed.

C.3. Frontier of Classifier Accuarcy against KL Divergence

We additionally calculate the accuracy of the sentiment classifier by taking the sigmoid of the oracle reward and plot the frontier of the accuracy vs reverse KL in Figure 7.

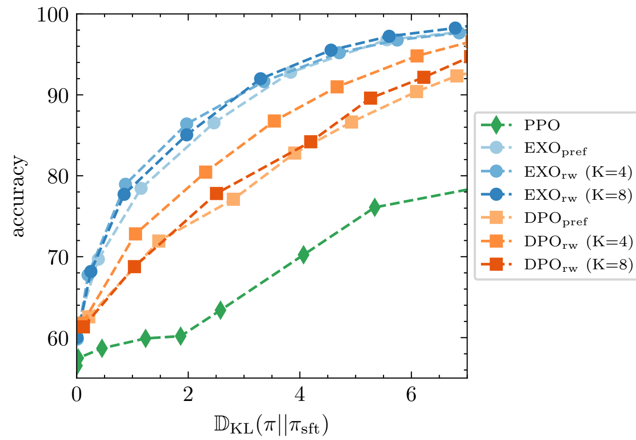


Figure 7. The frontier of classifier accuracy vs reverse KL to the SFT policy of different methods in the controlled experiment.

C.4. Optimization Efficiency

To demonstrate the efficency of optimizing the oracle reward, we plot the curve of the oracle reward with standard deviation on the test set of different training steps. Specifically, the standard deviation is calculated on the 4 samples generated given the same prompt and then averaged across 512 prompts.

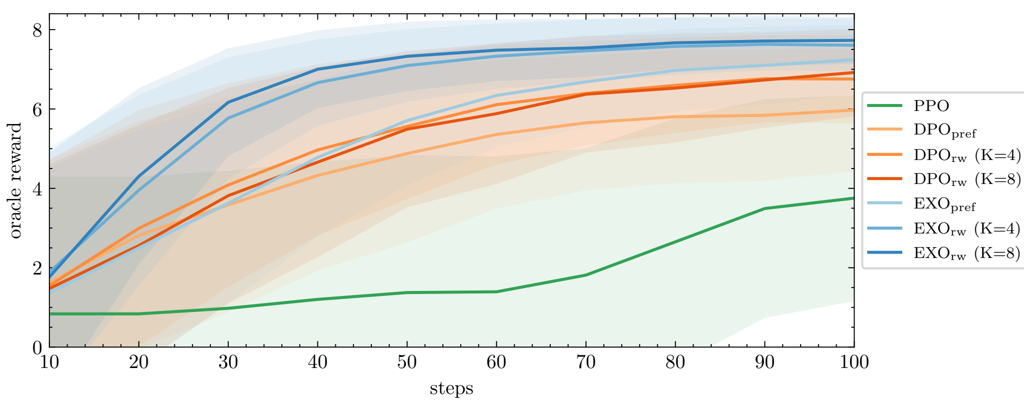


Figure 8. The curve of the oracle reward with standard deviation on the test set versus the training steps of different methods in the controlled experiment.

C.5. Additional Visualization Results

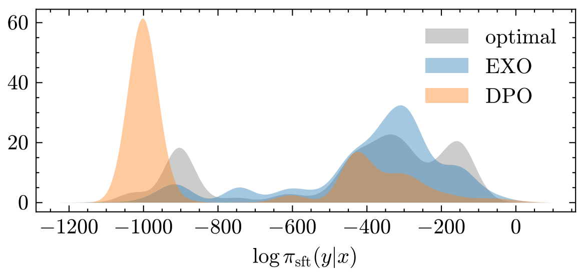
We additionally sample more prompts from the test set of the IMDb dataset, and plot the estimated density ratio of the policy learned by EXO and DPO and the optimal policy in Figure 9. Similar results were obtained among these diverse set of prompts: the density ratio under the EXO policy better matches the optimal policy at the high probability region against π_{sft} than the DPO policy. And very appearently, the DPO policy tends to overestimate the long tail in π_{sft} and thus often loses resolution in capturing the modes of the optimal policy.

C.6. Sample Outputs with GPT-4 Judgments

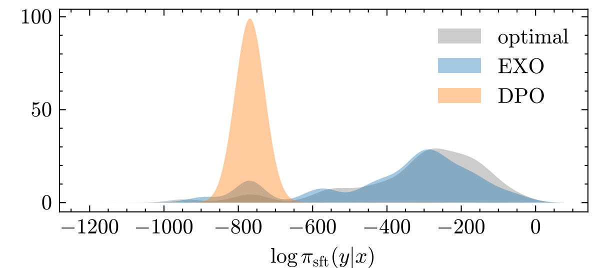
We present the sample outputs generated by EXO and DPO and the judgments made by GPT-4 when compared to the chosen response in the preference dataset. Table 6 and Table 7 show examples on the task of dialogue generation and summarization respectively.

Dialogue history
Human: How do I make a fishtail braid? Assistant: Hey human, just walk down the left and then down the right side, then at the end, you pull through. Are you looking to just braid, or a tight ponytail? Human: Yes. Can you tell me the steps to do it?
Chosen response
Assistant: Sure, here's how to do a side braid: 1) Pull hair up to the back of your head, leaving it out from the rest of your hair so you can see it. 2) Put one hair on the side where you want to make your braid. 3) Hold the braid with your thumb and forefinger 4) Bring your braid through the opposite side (putting the hair to the other side of the braid. 5) Repeat steps 2, 3, and 4 as needed until you've completed your braid.
EXO response
Assistant: Sure! I'll walk you through it step by step: 1. Stand with your left hand on the back of your head. Grab a left side section of hair, using the right hand to pull it through. 2. Take a right side section of hair, and pull it through 3. Grab a left side section of hair, and pull it through 4. Grab a right side section of hair, and pull it through 5. Repeat steps 2-4 for the last section of hair 6. Wrap the hair around your head and secure it at the nape of your neck.
GPT-4 Judgment (A = EXO, B = Chosen)
Response A provides a more detailed and step-by-step guide on how to make a fishtail braid, while Response B seems to describe a different type of braid and is less clear.
GPT-4 Judgment (A = Chosen, B = EXO)
Response B is more helpful because it provides a clearer, step-by-step guide on how to make a fishtail braid, whereas Response A seems to describe a different type of braid and is less clear.
DPO response
Assistant: Sure, let's walk through it. You'll want to start with a braid about two and a half inches long, then you want to braid it together and pull through, which means you'll somehow thread it through your own hair and tie it off at the other end. Then you're going to braid it again, but this time it will be a fishtail, since you'll pull it through. You'll keep doing this, braiding the fishtail.
GPT-4 Judgment (A = DPO, B = Chosen)
Response B is more helpful as it provides a clear, step-by-step guide on how to make a fishtail braid, while Response A is confusing and lacks clarity.
GPT-4 Judgment (A = Chosen, B = DPO)
Response A provides a clear, step-by-step guide on how to make a side braid, while Response B gives a vague description of how to make a fishtail braid without clear steps.

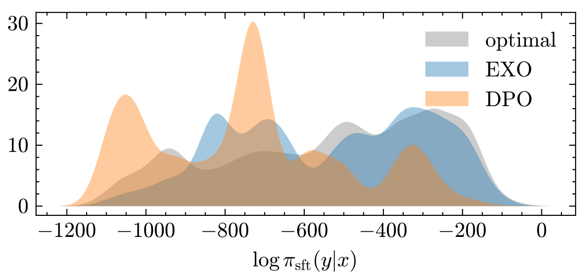
Table 6. Responses generated by EXO and DPO given the dialogue history sampled from the Anthropic-HH test set. GPT-4 consistently prefers EXO and disprefers DPO over the chosen response regardless of the order in which the evaluated pairs are presented.



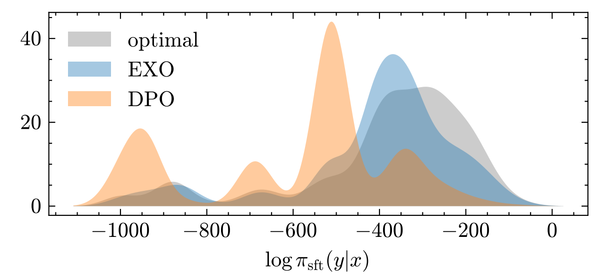
Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*Is this supposed to be serious? I hope not*”.



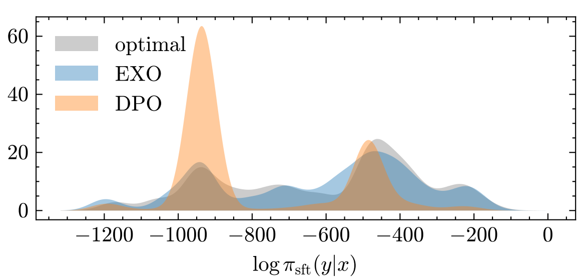
Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*Great book, great movie, great soundtrack. Frank*”.



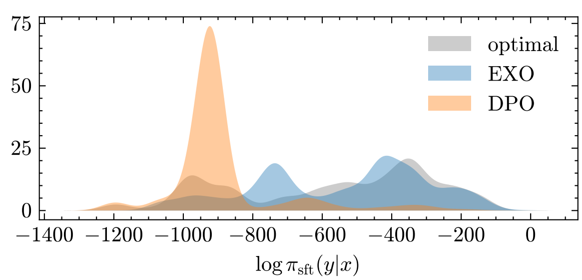
Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*This is indeed the film that popularized kung*”.



Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*This movie is about a group of people who are*”.



Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*What we have here the standard Disney direct to DVD*”.



Estimated density ratio of the EXO, DPO and optimal policy given the prompt “*Once the slow beginning gets underway, the film kicks*”.

Figure 9. Visualization of the estimated density ratio of the policy learned by EXO and DPO, and the optimal policy given 10 test prompts randomly sampled from the IMDb dataset.

Post
SUBREDDIT: r/Dogtraining TITLE: Best walks when on a choke chain: dog isn't being choked, but is it bringing up bad memories? POST: My 3 year old GSD x Husky rescue stayed with some relatives for a few days while my boyfriend and I moved house, and the relatives found an old choke chain to walk him with, instead of his Mekuti harness. We have had real problems with him pulling on the lead, and the Mekuti harness has worked wonders, but didn't eliminate the problem. He hasn't pulled when on the choke chain at all, even when it was first put on apparently, so we think that is what he was trained on (and damage from that process is probably why his barks and whines sound a bit odd). We now think the people who had him before us were violent towards him (if you mime kicking he runs across the room terrified), so, my question is: he is best behaved on walks when wearing a choke chain, and it isn't hurting him at all, but will the harm chokers and the people using them caused him in the past affect him in any way now? I 100% DO NOT want to use a choke chain on him and I don't think anyone should ever, ever train a dog that way, but this chain is the only lead we've used that has completely solved the problem with him pulling, and he hasn't seemed stressed out by it. If he ever started pulling on the lead so that the choker was tight, I would stop using it straight away. The last thing I want is my dog being hurt. If I'd known these relatives were going to put one on him, I wouldn't have let them look after him. Still, it works and doesn't hurt him, my only worry is it'll do psychological damage somehow.
Chosen summary
TL;DR: 3yr old husky rescue pulling on choke chain w/ no damage, but using it on walks occasionally causes him stress. Can this damage actually cause harm in the future?
EXO summary
TL;DR: Best walks are on a choke chain but dog isn't pulling on lead, has had no problems with pull before, not sure if past violent treatment will affect him now, does it cause harm to dogs being on choke chains/when being walked on them?
GPT-4 Judgment (A = EXO, B = Chosen)
Summary A provides a more comprehensive and accurate portrayal of the post's points, despite a few misplaced details, whereas Summary B incorrectly states the dog is showing stress and pulling on the choke chain.
GPT-4 Judgment (A = Chosen, B = EXO)
Summary B is preferred because it more accurately summarizes the main points of the user's post by mentioning the choke chain's effect on walks, mentioning the absence of negative effects of pull so far, and raising concerns about past violent treatment potentially affecting the dog's behavior.
DPO summary
TL;DR: If you took a dog and had it wear chokes it's self-choked, will it be traumatized in any way when you put it on and no longer train it to be choked?
GPT-4 Judgment (A = DPO, B = Chosen)
Summary B is preferred as it captures the key details more accurately, including the dog's breed, the issue of pulling, and the potential future harm, while Summary A is ambiguous and loss a lot of context.
GPT-4 Judgment (A = Chosen, B = DPO)
Summary A is more accurate and detailed because it correctly describes the specific situation of the dog and the concerns associated with using a choke chain, while Summary B is vague and confusing in its wording.

Table 7. Summaries generated by EXO and DPO given the post sampled from the TL;DR test set. GPT-4 consistently prefers EXO and disprefers DPO over the chosen summary regardless of the order in which the evaluated pairs are presented.