Token-level Direct Preference Optimization

Yongcheng Zeng [1 2] **Guoqing Liu** [3] **Weiyu Ma** [1 2] **Ning Yang** [1] **Haifeng Zhang** [1] **Jun Wang** [4]

Abstract

Fine-tuning pre-trained Large Language Models (LLMs) is essential to align them with human values and intentions. This process often utilizes methods like pairwise comparisons and KL divergence against a reference LLM, focusing on the evaluation of full answers generated by the models. However, the generation of these responses occurs in a token level, following a sequential, auto-regressive fashion. In this paper, we introduce Token-level Direct Preference Optimization (TDPO), a novel approach to align LLMs with human preferences by optimizing policy at the token level. Unlike previous methods, which face challenges in divergence efficiency, TDPO incorporates forward KL divergence constraints for each token, improving alignment and diversity. Utilizing the Bradley-Terry model for a token-based reward system, TDPO enhances the regulation of KL divergence, while preserving simplicity without the need for explicit reward modeling. Experimental results across various text tasks demonstrate TDPO's superior performance in balancing alignment with generation diversity. Notably, fine-tuning with TDPO strikes a better balance than DPO in the controlled sentiment generation and single-turn dialogue datasets, and significantly improves the quality of generated responses compared to both DPO and PPO-based RLHF methods. Our code is open-sourced at https://github.com/Vance0124/Token-level-Direct-Preference-Optimization.

---
[1]Institute of Automation, Chinese Academy of Sciences [2]School of Artificial Intelligence, University of Chinese Academy of Sciences [3]Microsoft Research AI4Science [4]University College London. Correspondence to: Jun Wang <jun.wang@cs.ucl.ac.uk>, Haifeng Zhang <haifeng.zhang@ia.ac.cn>.

1. Introduction

Large language models (LLMs) (Achiam et al., 2023; Bubeck et al., 2023) have demonstrated significant generalization capabilities in various domains including text summarization (Stiennon et al., 2022; Koh et al., 2022), coding writing (Chen et al., 2021; Gao et al., 2023), and even following human instructions (Chung et al., 2022; Ouyang et al., 2022). In order to align LLMs with human intentions, Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Dong et al., 2023; Yuan et al., 2023; Liu et al., 2023) has emerged as a highly effective method, embodying both stylistic and ethical values (Bai et al., 2022; Ganguli et al., 2022). These approaches typically involve the training of a reward model followed by the fine-tuning of the policy model using reinforcement learning (RL).

Direct Preference Optimization (DPO) (Rafailov et al., 2023) introduces a straightforward and effective technique for training LLMs using pairwise comparisons, without the need for explicitly establishing a reward model. DPO utilizes KL divergence to ensure that the training process remains closely aligned with a reference Large Language Model (LLM), preventing significant deviations. In DPO, KL divergence is assessed at the sentence level, reflecting the fact that evaluations are based on complete responses (answers), typically comprising several sentences. However, the generation of these responses occurs sequentially, following an auto-regressive approach. A potential benefit is to examine divergence in relation to a reference LLM on a more granular, token-by-token basis. One approach involves using sequential KL divergence (as defined in Definition 4.3), which monitors the trajectory of the generated responses. As illustrated in Figure 1, DPO demonstrates a significantly faster increase in KL divergence within the subset of less preferred responses when compared to the subset that is preferred. This results in an expanding gap between the two subsets and also indicates that DPO does not effectively control the KL divergence of the dispreferred response subset. This impacts the model's divergence efficiency and ultimately affects its linguistic capabilities and generative diversity. Such a limitation highlights the decreased effectiveness of employing KL divergence within the DPO framework, suggesting an area for improvement in its methodology.
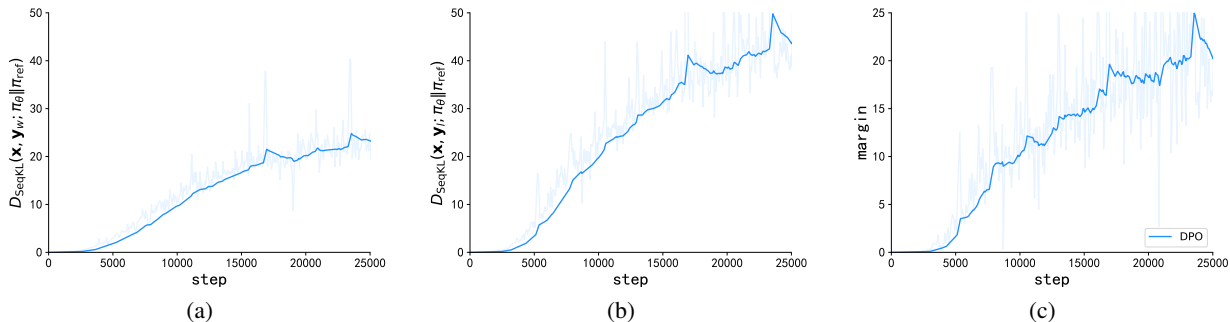
Figure 1. Sequential KL (SeqKL) divergence of both preferred response and dispreferred responses on IMDb dataset. Figure 1(a) shows the progression of SeqKL divergence on the preferred responses over training steps. Figure 1(b) depicts the evolution of SeqKL divergence on the dispreferred responses over the training steps. Figure 1(c) illustrates the difference between the SeqKL divergence of the dispreferred responses and that of the preferred responses during the training process, namely $margin = |D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta) - D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)|$. The definition of SeqKL divergence refers to Definition 4.3.

The imbalance in the growth rates of the sequential KL divergence is potentially related to the reverse KL divergence constraint employed by DPO. The mode-seeking property of reverse KL divergence tends to induce diversity reduction during generation, limiting the model's potential to produce diverse and effective responses (Wiher et al., 2022; Khalifa et al., 2020; Glaese et al., 2022; Perez et al., 2022). Built upon DPO, the f-DPO method (Wang et al., 2023) studies the trade-off between alignment performance and generation diversity of LLMs under different divergence constraints. It highlights the advantages of the mass-covering behavior of forward KL divergence in enhancing model diversity and explores the impact of different divergence constraints. Nevertheless, f-DPO only independently discusses the changes in model behavior under either the reverse KL divergence or the forward KL divergence constraints. Essentially, it does not fundamentally enhance the DPO algorithm itself but rather strikes a balance between alignment performance and generating diversity by simply swapping different KL divergence constraints.

Inspired by the aforementioned observations, we define and examine the problem of aligning with human preferences from a sequential and token-level standpoint. Some concurrent work has also been conducted in this direction (Rafailov et al., 2024; Zhong et al., 2024). We introduce a new method, referred to as Token-level Direct Preference Optimization (TDPO), which aims to strike a better balance between alignment performance and generation diversity by controlling the KL divergence for each token. In order to achieve this, we redefine the objective of maximising restricted rewards in a sequential manner. The connection between sentence-level reward and token-level generation is established by the use of the Bellman equation. Afterwards, the Bradley-Terry model (Bradley & Terry, 1952) is

converted into a representation at the token level, demonstrating its close relationship with the Regret Preference Model (Knox et al., 2022; 2023). By utilizing this method, we effectively integrate forward KL divergence restrictions for each token in the final objective function, resulting in improved regulation of KL divergence.

TDPO maintains the simplicity of DPO while offering improved regulation of KL divergence for aligning LLMs with human preferences. Echoing the strategy of DPO, our method directly optimizes the policy without necessitating explicit reward model learning or policy sampling throughout the training phase. Our experimental results demonstrate the effectiveness of TDPO across multiple text tasks, and gain a notable enhancement in the quality of generated responses in comparison to both DPO and PPO-based RLHF methods. In conclusion, TDPO stands out for its ability to not only effectively address the issue of excessive KL divergence but also greatly improve divergence efficiency.

2. Related Works

The emergence of ChatGPT has catalyzed significant advancements in the field of Large Language Models (LLMs), such as OpenAI's GPT-4 (Achiam et al., 2023), Mistral (Jiang et al., 2023), and Google's Gemini (Team et al., 2023). Generally, the training of LLMs involves three stages: initial unsupervised pre-training on massive text corpora to grasp linguistic structures (Raffel et al., 2020; Brown et al., 2020; Workshop et al., 2022; Touvron et al., 2023), followed by supervised fine-tuning with task-specific datasets to enhance the LLMs' probability of producing desired responses (Taori et al., 2023; Chiang et al., 2023; Vu et al., 2023). However, due to the typically limited and expensive availability of labeled datasets during the supervised

fine-tuning stage, the model may retain biases and inaccuracies, manifesting as societal biases (Sheng et al., 2021), ethical concerns (Weidinger et al., 2021), toxicity (Rauh et al., 2022), and hallucinations (Huang et al., 2023), which necessitates a subsequent AI alignment phase. Noteworthy models achieving significant alignment, such as Zephyr (Tunstall et al., 2023) and GPT-4 (Achiam et al., 2023), have demonstrated the effectiveness of techniques like Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) algorithms.

Reinforcement Learning from Human Feedback (RLHF) has emerged as a cornerstone in aligning LLMs with human values, providing a mechanism to refine model outputs based on qualitative feedback (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Song et al., 2023; Touvron et al., 2023). This approach has shown considerable promise in making models more responsive to human expectations and ethical considerations by iteratively improving their performance through human-generated feedback. However, the complexity of implementing RLHF, compounded by the inaccuracies in human-generated reward models (Wu et al., 2023), has prompted the exploration of alternative strategies. Methods like Reward Ranked Fine-Tuning (RAFT) (Dong et al., 2023) and Rank Responses to align Human Feedback (RRHF) (Yuan et al., 2023) offer streamlined approaches to alignment, circumventing some of RLHF's inherent challenges. Particularly, Direct Preference Optimization (DPO) (Rafailov et al., 2023) represents a breakthrough in direct policy optimization, addressing the intricacies of balancing model behavior through a nuanced approach to reward function optimization. Nevertheless, the challenge of maintaining linguistic diversity while aligning with human preferences remains a pivotal concern, prompting our proposed Token-level Direct Preference Optimization (TDPO), which seeks to harmonize the dual objectives of alignment accuracy and expressive range in model outputs.

3. Preliminaries

For language generation, a language model (LM) is prompted with prompt (question) x to generate a response (answer) y, where both x and y consist of a sequence of tokens. Direct Preference Optimization (DPO) (Rafailov et al., 2023) commences with the RL objective from the RLHF:

$$
\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \big[r(x, y)
$$
$$
- \beta D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x) \big\| \pi_{\mathrm{ref}}(\cdot \mid x)\big)\big], \tag{1}
$$

where $\mathcal{D}$ represents the human preference dataset, $r(x, y)$ denotes the reward function, $\pi_{\mathrm{ref}}(\cdot|x)$ serves as a reference model, typically chosen the language model after supervised fine-tuning, π_θ represents the model undergoing RL fine-

tuning, initialized with $\pi_\theta = \pi_{\mathrm{ref}}$, and β is the coefficient for the reverse KL divergence penalty.

By directly deriving from Eq. 1, DPO establishes a mapping between the reward model and the optimal policy under the reverse KL divergence, obtaining a representation of the reward function concerning the policy:

$$
r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)} + \beta \log Z(x). \tag{2}
$$

Here, $Z(x)$ is the partition function.

To align with human preference, DPO uses the Bradley-Terry model for pairwise comparisons:

$$
P_{\mathrm{BT}}(y_1 \succ y_2|x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}. \tag{3}
$$

By substituting Eq. 2 into Eq. 3 and leveraging the negative log-likelihood loss, DPO derives the objective function:

$$
u(x, y_w, y_l) = \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)},
$$
$$
\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(u(x, y_w, y_l)\right)\right], \tag{4}
$$

and the derivative is given as follows:

$$
\nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[(-u)\nabla_\theta u\right], \tag{5}
$$

where u is the abbreviation of $u(x, y_w, y_l)$, y_w and y_l denotes the preferred and dispreferred completion.

4. Methodology

In this section, we initially reformulate the constrained reward maximization problem into a token-level form. From this, we derive the mapping between the state-action function and the optimal policy. Subsequently, we convert the Bradley-Terry model into token-level representation, establishing its equivalence with the Regret Preference Model. By substituting the mapping relationship into the reward model in token-level format, we obtain the optimization objective solely related to the policy. Finally, we conduct a formalized analysis of this optimization objective in terms of derivatives and, based on this, derive the ultimate loss function for TDPO.

4.1. Markov Decision Process under Token Rewards

To model the sequential, auto-regressive generation, we extend the sentence-level formulation in Section 3 by considering that the response consists of T tokens $y = y^{<T+1} :=$ $[y^1, y^2, ..., y^T]$, where $y^t \in \mathcal{Y}$, and $\mathcal{Y}$ represents the alphabet (vocabulary). Additionally, we assume $y^{<1} = [\ ]$. Given a prompt x and the first $t - 1$ tokens $y^{<t}$ of the response

y, the LM predicts the probability distribution of the next token $\pi_\theta(\cdot|[x, y^{<t}])$.

When modeling text generation as a Markov decision process (Puterman, 2014), a state is a combination of the prompt and the generated response up to the current step, denoted as $s_t = [x, y^{<t}]$. An action corresponds to the next generated token, denoted as $a_t = y^t$, and the token-wise reward is defined as $R_t := R(s_t, a_t) = R([x, y^{<t}], y^t)$.

Expanding on the provided definitions, we establish the state-action function Q_π, the state value function V_π and the advantage function A_π for a policy π:

$$
\begin{aligned}
Q_\pi([x, y^{<t}], y^t) &= \mathbb{E}_\pi\left[\sum_{k=0}^\infty \gamma^k R_{t+k}\Big|s_t = [x, y^{<t}], a_t = y^t\right], \\
V_\pi([x, y^{<t}]) &= \mathbb{E}_\pi[Q_\pi([x, y^{<t}], y^t)|s_t = [x, y^{<t}]], \\
A_\pi([x, y^{<t}], y^t) &= Q_\pi([x, y^{<t}], y^t) - V_\pi([x, y^{<t}]).
\end{aligned}
\tag{6}
$$

where γ represents the discount factor. In this paper, we set $\gamma = 1$.

4.2. Token-Level Optimization

DPO's objective function in Eq. 1 operates at the sentence level. In contrast, we propose an alternative token-level objective function:

$$
\begin{aligned}
\max_{\pi_\theta} \mathbb{E}_{x,y^{<t}\sim\mathcal{D}, z\sim\pi_\theta(\cdot|[x,y^{<t}])} \big[&A_{\pi_{\text{ref}}}([x, y^{<t}], z) \\
&- \beta D_{\text{KL}}\left(\pi_\theta(\cdot|[x, y^{<t}])||\pi_{\text{ref}}(\cdot|[x, y^{<t}])\right)\big].
\end{aligned}
\tag{7}
$$

The objective function is inspired by Trust Region Policy Optimization (TRPO) (Schulman et al., 2015). As demonstrated in Lemma 4.1, maximizing the objective function in Eq. 7 will result in policy improvements in terms of expected return.

Lemma 4.1. *Given two policies π and $\tilde{\pi}$, if for any state $s_t = [x, y^{<t}]$, $\mathbb{E}_{z\sim\tilde{\pi}}[A_\pi([x, y^{<t}], z)] \geq 0$, then we can conclude:*

$$
\mathbb{E}_{x\sim\mathcal{D}}[V_{\tilde{\pi}}([x])] \geq \mathbb{E}_{x\sim\mathcal{D}}[V_\pi([x])],
$$

The proof is provided in Appendix A.1.

Notably, to maintain generation diversity and prevent the model from hacking some high-reward answers, we incorporate reverse KL divergence for each token in our token-level objective function, which prevents the model from deviating too far from the reference model distribution.

Starting from the token-level objective function in Eq. 7, we can directly derive the mapping between the state-action function Q_π and the optimal policy π_θ^*. We summarize this relationship in the following lemma.

Lemma 4.2. *The constrained problem in Eq. 7 has the closed-form solution:*

$$
\begin{aligned}
\pi_\theta^*(z|[x, y^{<t}]) = \\
\frac{\pi_{\text{ref}}(z|[x, y^{<t}]) \exp\left(\frac{1}{\beta}Q_{\pi_{\text{ref}}}([x, y^{<t}], z)\right)}{Z([x, y^{<t}]; \beta)},
\end{aligned}
\tag{8}
$$

where $Z([x, y^{<t}]; \beta) = \mathbb{E}_{z\sim\pi_{\text{ref}}(\cdot|[x,y^{<t}])}e^{\frac{1}{\beta}Q_{\pi_{\text{ref}}}([x,y^{<t}],z)}$ is the partition function.

See Appendix A.2 for more details.

To obtain the optimal policy π_θ^* from Eq. 8, we must estimate the state-action function $Q_{\pi_{\text{ref}}}$ and the partition function $Z(\cdot)$. However, ensuring the accuracy of the state-action function Q_π at each state and action is challenging, and estimating the partition function $Z(\cdot)$ is also difficult. Therefore, we reorganize Eq. 8 to obtain the expression of the state-value function in terms of the policy:

$$
\begin{aligned}
Q_{\pi_{\text{ref}}}([x, y^{<t}], z) = \\
\beta \log \frac{\pi_\theta^*(z|[x, y^{<t}])}{\pi_{\text{ref}}(z|[x, y^{<t}])} + \beta \log Z([x, y^{<t}]; \beta).
\end{aligned}
\tag{9}
$$

4.3. BT Model Reformulation via Advantage Function

To facilitate subsequent derivations, we first introduce the sequential KL divergence, as defined in Definition 4.3.

Definition 4.3. Given two language models π_1 and π_2, with the input prompt x and output response y, the sequential KL divergence is defined as:

$$
\begin{aligned}
D_{\text{SeqKL}}(x, y; \pi_1\|\pi_2) = \\
\sum_{t=1}^T D_{\text{KL}}(\pi_1(\cdot|[x, y^{<t}])\|\pi_2(\cdot|[x, y^{<t}])).
\end{aligned}
\tag{10}
$$

Given prompts x and pairwise responses (y_1, y_2), the Bradley-Terry model expresses the human preference probability. However, since the Bradley-Terry model is formulated at the sentence level, it cannot establish a connection with the token-level mapping presented in Eq. 9. Consequently, we need to derive a token-level preference model. Initiating from the Bradley-Terry model, we transform it into a token-level formulation and demonstrate its equivalence with the Regret Preference Model (Knox et al., 2023; 2022), as shown in the Lemma 4.4.

Lemma 4.4. *Given a reward function $r(\text{x}, \text{y})$, assuming a relationship between token-wise rewards and the reward function represented by $r(x, y) = \sum_{t=1}^T \gamma^{t-1} R([x, y^{<t}], y^t)$, we can establish the equivalence between the Bradley-Terry model and the Regret Pref-*

erence Model in the task of text generation alignment, i.e.,

$$P_{\text{BT}}(y_1 \succ y_2|x) =$$

$$\sigma \left(\sum_{t=1}^{T_1} \gamma^{t-1} A_\pi([x, y_1^{<t}], y_1^t) - \sum_{t=1}^{T_2} \gamma^{t-1} A_\pi([x, y_2^{<t}], y_2^t) \right),$$

(11)

where $\sigma(x) = 1/(1 + exp(-x))$ is the logistic sigmoid function.

We prove this lemma in A.3.

In Lemma 4.4, we assume that $r(x, y) = \sum_{t=1}^{T} \gamma^{t-1} R([x, y^{<t}], y^t)$. This assumption is natural in the context of RL, where $r(x, y)$ represents the overall reward for response y given the prompt x. Considering text generation as a sequential decision-making problem, $r(x, y)$ can be viewed as the cumulative reward for the generated text.

According to the definition of the advantage function in Section 4.1, we can directly establish the relationship between the optimal solution in Eq. 9 and preference optimization objective in Eq. 11. One intractable aspect is that the state-action function Q_π depends on a partition function, which is contingent on both the input prompt x and the output response y. This results in non-identical values of the partition function for a pair of responses (y_w, y_l), specifically, $Z([x, y_w^{<t}]; \beta) \neq Z([x, y_l^{<t}]; \beta)$. As a result, we cannot employ a cancellation strategy similar to DPO, which relies on the property that the Bradley-Terry model depends only on the difference in rewards between two completions.

Fortunately, by expanding the advantage function A_π and converting the state-value function V_π into a form exclusively related to the state-action function Q_π, we can offset the partition function naturally. In this way, we ultimately reformulate the Bradley-Terry model to be directly tied to the optimal policy π_θ^* and the reference policy π_{ref}. This is summarized in the following theorem.

Theorem 4.5. *In the KL-constrained advantage function maximization problem corresponding to Eq.7, the Bradley-Terry model express the human preference probability in terms of the optimal policy π_θ^* and reference policy π_{ref}:*

$$P_{\text{BT}}^*(y_1 \succ y_2|x) = \sigma(u^*(x, y_1, y_2) - \delta^*(x, y_1, y_2)), \quad (12)$$

where, $u(x, y_1, y_2)$ refers to the difference in rewards implicitly defined by the language model π_θ and the reference model π_{ref} (Rafailov et al., 2023), represented as

$$u(x, y_1, y_2) = \beta \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}, \quad (13)$$

and $\delta(x, y_1, y_2)$ refers to the difference in sequential forward KL divergence between two pairs (x, y_1) and (x, y_2), weighted by β, expressed as

$$\delta(x, y_1, y_2) = \beta D_{\text{SeqKL}}(x, y_2; \pi_{\text{ref}} \| \pi_\theta)$$
$$- \beta D_{\text{SeqKL}}(x, y_1; \pi_{\text{ref}} \| \pi_\theta). \quad (14)$$

The proof is provided in the Appendix A.4.

4.4. Loss Function and Formal Analysis

Drawing on Eq. 12, we reformulate the Bradley-Terry model into a structure solely relevant to the policy. This allows us to formulate a likelihood maximization objective for a parametrized policy π_θ, leading to the derivation of the loss function for the initial version of our method, TDPO$_1$:

$$\mathcal{L}_{\text{TDPO}_1}(\pi_\theta; \pi_{\text{ref}}) =$$
$$- \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(u(x, y_w, y_l) - \delta(x, y_w, y_l) \right) \right].$$
(15)

Through this approach, we explicitly introduce sequential forward KL divergence into the loss function. Coupled with the implicitly integrated reverse KL divergence, we enhance our ability to balance alignment performance and generation diversity of LLMs.

Subsequently, we conduct a derivative analysis of our method and make specific modifications to the loss function of TDPO. For convenience, we use u to denote $u(x, y_w, y_l)$, and δ to represent $\delta(x, y_w, y_l)$. By employing the formulation of the loss function presented in Eq.15, we compute the gradient of the loss function with respect to the parameters θ:

$$\nabla_\theta \mathcal{L}_{\text{TDPO}_1}(\pi_\theta; \pi_{\text{ref}}) =$$
$$- \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[(-u + \delta) \left[\nabla_\theta u - \nabla_\theta \delta \right] \right].$$
(16)

In Eq. 16, $(-u + \delta)$ serves as the weighting factor for the gradient. The first part $(-u)$ corresponds to the weight factor in the loss function of DPO. When the language model makes errors in predicting human preferences, i.e., $\log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} > \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}$, the value of $(-u)$ will become larger, applying a stronger update for the comparison (y_w, y_l). While the second part δ is a distinctive component of our method. As shown in Figure 1, the KL divergence growth rate for the dispreferred response subset is faster than that for the preferred response subset. With the increasing disparity, the corresponding value of δ rises, thereby amplifying the weight factor $(-u + \delta)$. Combined with the subsequent gradient term, our objective function can effectively suppress the difference in KL divergence between pairs of responses with large disparities in KL divergence. Through the collaborative influence of the weight factor δ and the gradient term $(-\nabla_\theta \delta)$, our method achieves the purpose of automatic control over the KL divergence balance.

The gradient of the loss function in Eq. 16 also consists of two components, $\nabla_\theta u$ and $(-\nabla_\theta \delta)$. $\nabla_\theta u$ represents the optimization direction of the gradient in DPO. Intuitively, $\nabla_\theta u$ increases the likelihood of preferred completions y_w and decreases the likelihood of dispreferred com-

pletions y_l. While $(-\nabla_\theta \delta)$ tends to narrow the gap between $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta)$ and $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)$.

However, when considered separately, the gradient of $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} | \pi_\theta)$ in the loss function tends to increase the sequential KL divergence between π_{ref} and π_θ at (x, y_w) during the optimization process. This is because the sequential forward KL divergence in the loss function is introduced through the state-value function V_π, inherently introducing an expectation $\mathbb{E}_{z \sim \pi_{\text{ref}}} \left[\log \frac{\pi_\theta(z|[x, y^{<t}])}{\pi_{\text{ref}}(z|[x, y^{<t}])} \right]$ as a baseline at each token. The negative value of this expectation corresponds precisely to a forward KL divergence $D_{\text{KL}}(\pi_{\text{ref}}(\cdot|[x, y^{<t}]) | \pi_\theta(\cdot|[x, y^{<t}]))$, which can be used to constrain the unbalanced growth of KL divergence. For the prompt x and the preferred response y_w, at each token, the loss function in Eq. 16 tends to increase the likelihood of $\log \frac{\pi(y_w^t|[x, y_w^{<t}])}{\pi_{\text{ref}}(y_w^t|[x, y_w^{<t}])}$ while simultaneously decreasing the expectation, enlarging the gap between the specified term y_w^t and the baseline to expedite training. The impact of decreasing the expectation is an increase in the forward KL divergence $D_{\text{KL}}(\pi_{\text{ref}}(\cdot|[x, y_w^{<t}]) | \pi_\theta(\cdot|[x, y_w^{<t}]))$ at each token, leading to an increase in $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} | \pi_\theta)$. As we do not aim to accelerate the training speed and prefer to ensure training stability, we modify the loss function by discontinuing the gradient propagation of $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} | \pi_\theta)$ and treating it as a baseline term for alignment of $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} | \pi_\theta)$.

Different from $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} | \pi_\theta)$, the gradient of $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} | \pi_\theta)$ tends to reduce the sequential KL divergence between π_{ref} and π_θ at (x, y_l). For the prompt x and the rejected response y_l, the loss function in Eq.16 tends to decrease the likelihood of $\log \frac{\pi(y_l^t|[x, y_l^{<t}])}{\pi_{\text{ref}}(y_l^t|[x, y_l^{<t}])}$ at each token while increasing the expectation $\mathbb{E}_{z \sim \pi_{\text{ref}}} \left[\log \frac{\pi_\theta(z|[x, y_l^{<t}])}{\pi_{\text{ref}}(z|[x, y_l^{<t}])} \right]$. The increase in the expectation implies a smaller forward KL divergence at that token, thereby acting to constrain the growth rate of sequential forward KL divergence. Therefore, for this term, we choose to retain its gradient updates.

In conclusion, we only propagate the gradient of the $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} | \pi_\theta)$ in $(-\nabla_\theta \delta)$. When the second part weight factor δ becomes larger, it imposes a stronger suppression on $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)$ to control the balance of KL divergence.

Furthermore, to achieve a better balance between alignment performance and generation diversity in TDPO, we introduce an additional parameter α into the loss function. By adjusting the magnitude of α, we can control the deviation between $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta)$ and $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)$.

In summary, we modify the loss function of TDPO_1, resulting in the second version of our method, TDPO_2, as

follows:

$$
\mathcal{L}_{\text{TDPO}_2}(\pi_\theta; \pi_{\text{ref}}) = \\
- \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(u(x, y_w, y_l) - \alpha \delta_2(x, y_w, y_l) \right) \right],
$$
(17)

where α is a parameter, and

$$
\delta_2(x, y_1, y_2) = \beta D_{\text{SeqKL}}(x, y_2; \pi_{\text{ref}} \| \pi_\theta) \\
- sg \left(\beta D_{\text{SeqKL}}(x, y_1; \pi_{\text{ref}} \| \pi_\theta) \right).
$$
(18)

The sg represents the stop-gradient operator, which blocks the propagation of gradients.

We summarize the comparison of the loss functions for DPO, TDPO_1, and TDPO_2, as presented in Figure 2.

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right]$$

$$\mathcal{L}_{\text{TDPO}_1}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log \sigma \left(\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) - \left(\beta D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta) - \beta D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta) \right) \right) \right]$$

$$\mathcal{L}_{\text{TDPO}_2}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log \sigma \left(\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) - \alpha \left(\beta D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta) - sg \left(\beta D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta) \right) \right) \right) \right]$$

Figure 2. Comparison of Loss Functions for DPO, TDPO_1 and TDPO_2 Methods. The sg denotes the stop-gradient operator. Both TDPO_1 and TDPO_2 incorporate an additional term for finer-grained control over the KL divergence, compared to DPO.

Leveraging the parameter β to regulate the deviation of the language model from the base reference model, and α to control the balance of sequential KL divergence within the language model, our approach achieves superior alignment with human preferences while preserving model generation diversity effectively. We provided the pseudocode in Algorithm 1 and the Pytorch implementation version of TDPO loss in Appendix B.

5. Experiments

In this section, we demonstrate the superior performance of our algorithm in three different open-sourced datasets: the IMDb sentiment dataset (Maas et al., 2011), the Anthropic HH dataset (Bai et al., 2022), and MT-bench (Zheng et al., 2023). The IMDb dataset serves as a controlled semantic generation dataset where the model is presented with prompts consisting of prefixes from movie reviews, and required to generate responses with positive sentiment. The Anthropic HH dataset is a single-turn dialogue dataset where the model receives human queries, covering various

---

Algorithm 1 Token-level Direct Preference Optimization (TDPO)

1: **Input:** Reference model π_{ref}, Policy model π_θ, Coefficient α, β, Learning rate η
2: **Input:** Dataset $\mathcal{D} = \{(x, y_w, y_l)^i\}_{i=1}^N$ of size N, Method $\mathcal{M}$
3: **Initialize:** $\pi_\theta \leftarrow \pi_{\text{ref}}$
4: **for** each epoch **do**
5:     Sample mini-batch $\mathcal{D}_m = \{(x, y_w, y_l)^m\}_{m=1}^M$ from $\mathcal{D}$
6:     Predict the probabilities $\pi_\theta(y_w|x)$ and $\pi_\theta(y_l|x)$ for (x, y_w, y_l) in the mini-batch $\mathcal{D}_m$ using the policy model
7:     Predict the probabilities $\pi_{\text{ref}}(y_w|x)$ and $\pi_{\text{ref}}(y_l|x)$ for (x, y_w, y_l) in the mini-batch $\mathcal{D}_m$ using the reference model
8:     Calculate the function $u(x, y_w, y_l) = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$ ▷ Eq.13
9:     Compute the sequential KL divergence $D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta)$ for (x, y_w) in the mini-batch $\mathcal{D}_m$
10:    Compute the sequential KL divergence $D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)$ for (x, y_l) in the mini-batch $\mathcal{D}_m$
11:    **if** Method $\mathcal{M}$ is TDPO$_1$ **then**
12:       Calculate the function $\delta(x, y_w, y_l) = \beta D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta) - \beta D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta)$ ▷ Eq.14
13:       $\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_m} [\log \sigma (u(x, y_w, y_l) - \delta(x, y_w, y_l))]$ ▷ Eq.15
14:    **else if** Method $\mathcal{M}$ is TDPO$_2$ **then**
15:       Calculate the function $\delta_2(x, y_w, y_l) = \beta D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta) - sg(\beta D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta))$ ▷ Eq.18
16:       $\theta \leftarrow \theta + \eta \nabla_\theta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_m} [\log \sigma (u(x, y_w, y_l) - \alpha \delta_2(x, y_w, y_l))]$ ▷ Eq.17
17:    **end if**
18: **end for**
19: **Output:** π_θ

---

topics such as academic questions or life guidance. The trained model is tasked with providing helpful answers to these questions while avoiding toxic responses. Finally, MT-Bench is a GPT-4-based evaluation benchmark, assessing the proficiency of LLMs in handling multi-turn open-ended questions. Questions in MT-Bench span eight distinct knowledge domains, from areas such as writing, mathematical reasoning, and humanities. Experimental results demonstrate that MT-Bench achieves consistency with human preferences exceeding 80%.

5.1. Experiments on IMDb Dataset

In this experiment, besides our proposed methods TDPO$_1$ and TDPO$_2$, we also implemented the DPO algorithm for fair comparison. We employed GPT-2 Large (Radford et al., 2019) as our base model and the model checkpoint: *insub/gpt2-large-IMDb-fine-tuned*[1] as the SFT model. During the evaluation, we utilized the pre-trained sentiment classifier *siebert/sentiment-roberta-large-english*[2] to compute rewards. For DPO, we followed the official implementation (Rafailov et al., 2023), setting β at 0.1. To analyze the effectiveness of each algorithm in optimizing the constrained reward maximization objective, we evaluated each algorithm after 100 training steps until convergence, computing its frontier of average reward and average sequential KL divergence with the reference policy.

---

[1]https://huggingface.co/insub/gpt2-large-IMDb-fine-tuned
[2]https://huggingface.co/siebert/sentiment-roberta-large-english

The results are depicted in Figure 3(a). We implement the DPO, TDPO$_1$, and different versions of TDPO$_2$ algorithms with varying the parameter α. From the figure, we notice that although DPO establishes an efficient frontier, TDPO$_1$ and TDPO$_2$ outperform DPO in terms of divergence versus reward on the frontier, achieving higher reward while maintaining low KL divergence. We also implemented versions of TDPO$_2$ with $\alpha \in \{1, 1.5, 2, 5\}$. However, we found that higher values of α made it difficult to optimize the reward. In Figures 3(b) to 3(d), we illustrate the curves portraying the sequential KL divergence for different algorithms during the training process. The sequential KL divergence growth rate of DPO on the dispreferred response subset is significantly higher than that on the preferred response subset, leading to an increasing offset between them. In contrast, TDPO$_2$ exhibits superior control over KL divergence, achieving better divergence efficiency compared to DPO. As analyzed in Section 4.4, TDPO$_1$ tends to result in an increased sequential KL divergence on the preferred response subset, thereby exhibiting a weaker capacity for KL divergence adjustment compared to TDPO$_2$. TDPO$_2$ maintains a more balanced sequential KL divergence on both dispreferred and preferred response subsets, contributing to its ability to achieve a superior frontier. Although a larger α enhances control over the sequential KL divergence, it also affects the speed and difficulty of optimization. For the remainder of this paper, we set $\alpha = 0.5$. In Appendix C, we also present graphs of the frontier between the reward and forward KL divergence and the progression curves of the forward KL divergence throughout the training process.
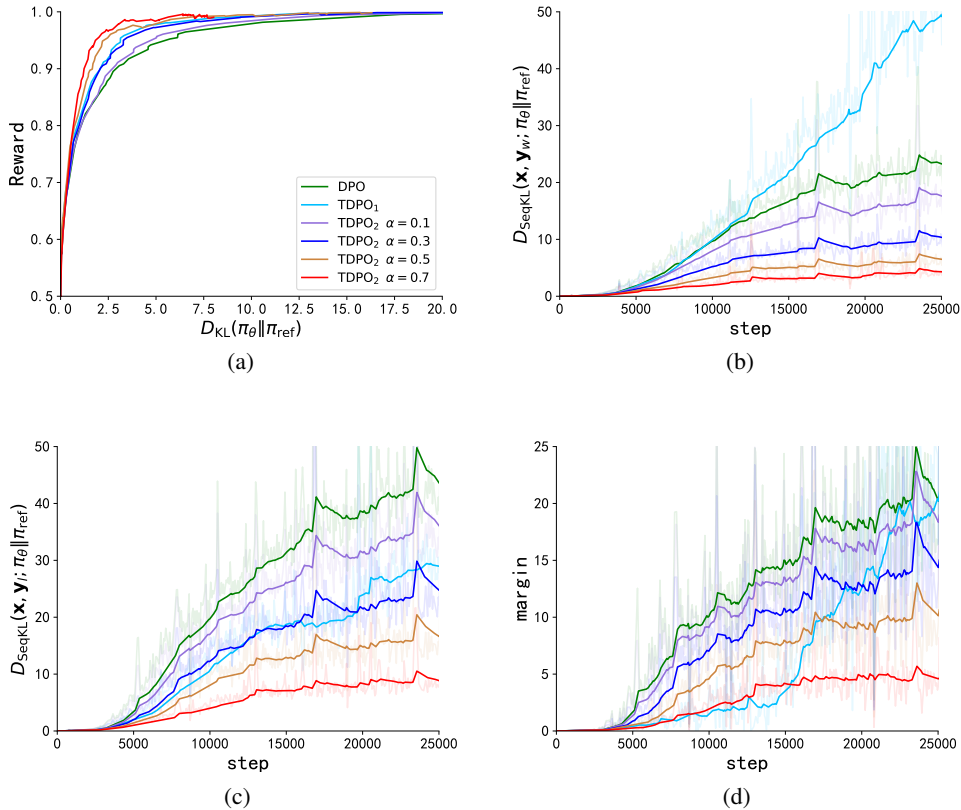
Figure 3. The experiment on IMDb dataset. Figure 3(a) represents the frontier of expected reward and KL divergence with respect to the reference model. We implemented DPO, TDPO_1, and different versions of TDPO_2 with respect to the parameter α. Both TDPO_1 and TDPO_2 outperform DPO in terms of the frontier, with TDPO_2 showing further improvement over TDPO_1. This demonstrates the effectiveness of our analysis and modifications. Figure 3(b) and Figure 3(c) present the progression of sequential KL divergence on the preferred and dispreferred responses subset over training steps respectively. Figure 3(d) illustrates the difference between the sequential KL divergence on the dispreferred responses subset and that on the preferred responses subset throughout the training process, namely $margin = |D_{\text{SeqKL}}(x, y_w; \pi_{\text{ref}} \| \pi_\theta) - D_{\text{SeqKL}}(x, y_l; \pi_{\text{ref}} \| \pi_\theta)|$. TDPO_2 exhibit superior regulation over KL divergence compared to the TDPO_1 and DPO algorithm.

Table 1. Comparison of DPO, TDPO_1 and TDPO_2 in terms of the trade-off between Alignment (accuracy) and Diversity (entropy) on the Anthropic HH dataset. The $\uparrow$ indicates higher values are preferable.

Method	Alignment Accuracy(%) $\uparrow$	Diversity Entropy $\uparrow$
f-DPO(FKL)	54.71	4.708
DPO	59.43	3.196
TDPO_1(**ours**)	60.08	4.727
TDPO_2(**ours**)	**67.33**	**4.915**

5.2. Experiments on Anthropic HH Dataset

Next, we evaluate the performance of TDPO_1 and TDPO_2 on the Anthropic HH dataset. We use Pythia-2.8B (Bider-

man et al., 2023) as the base model and fine-tune the base model on chosen completions to train a reference model, such that completions are within the distribution of the model. Subsequently, we train TDPO_1, TDPO_2, DPO (Rafailov et al., 2023) and f-DPO with forward KL divergence constraint (Wang et al., 2023) on this reference model. In this experiment, our primary focus is on two aspects: 1) the trade-off between alignment and diversity in generating responses among different algorithms, and 2) the ability of different algorithms to align with human preferences. For the first part, we utilize automatic metrics for evaluation, while for the second part, we rely on the GPT-4 evaluation. Both evaluations were conducted on the test set of the Anthropic HH dataset.

To assess the alignment performance of different algorithms in generating responses, we compute the accuracy of gen-

erated responses relative to chosen completions in the test dataset. To measure the diversity, we employ nucleus sampling with $p = 0.95$ to generate 25 responses and utilize the predictive entropy as the evaluation metric. The trade-off between alignment accuracy and diversity for different algorithms is summarized in Table 1. $TDPO_2$ not only surpasses DPO, f-DPO and $TDPO_1$ in terms of accuracy but also excels in entropy, achieving a superior balance between alignment and diversity.

To further assess the ability of $TDPO_1$ and $TDPO_2$ to align with human preferences, we evaluated the win rates of responses generated by models trained with different algorithms against chosen responses on the test set of the HH dataset, the result is illustrated in the Figure 4. Compared to the SFT model, the DPO, $TDPO_1$, and $TDPO_2$ algorithms better align with human preferences, achieving win rates not less than 50% against chosen responses at temperature 0.75. This demonstrates that both $TDPO_1$, and $TDPO_2$ possess a strong capability to align with human preferences.
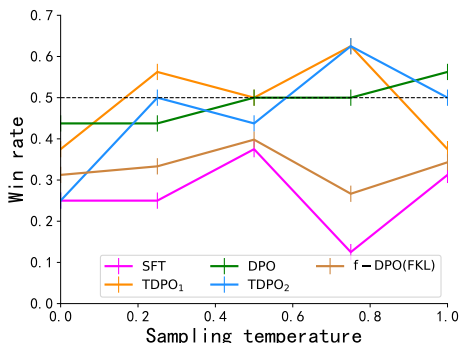

Figure 4. The win rates, computed by GPT-4, in comparison to the chosen responses for Anthropic-HH one-step dialogue.

5.3. Experiments on MT-Bench

To comprehensively evaluate $TDPO_1$, and $TDPO_2$ in terms of generation quality, we conducted pairwise comparisons on the MT-Bench using models trained on the Anthropic HH dataset. Following the official MT-Bench implementation, we sampled responses with a temperature coefficient of 0.7 and constrained the maximum number of newly generated tokens to 512. For the PPO baseline, we employed the trlx framework (Havrilla et al., 2023), utilizing the proxy reward model *Dahoas/gptj-rm-static*[3] during training. The result is depicted in the Figure 5. It reveals that $TDPO_2$ achieves a higher win rate compared to other algorithms, indicating its ability to assist LLMs in generating higher-quality responses. This advantage is attributed

---

[3] https://huggingface.co/Dahoas/gptj-rm-static

to its exceptional ability to regulate KL divergence, facilitating a better balance between performance alignment and generation diversity.
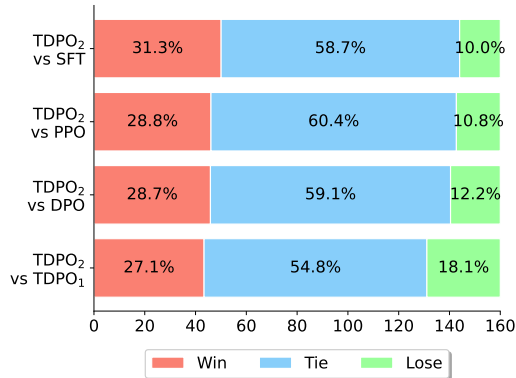

Figure 5. MT-Bench comparison between SFT, PPO, DPO, $TDPO_1$ and $TDPO_2$ methods. The win, tie, and lose rates are evaluated based on GPT-4.

6. Conclusion

In this work, we introduced Token-level Direct Preference Optimization (TDPO), an innovative token-level fine-tuning approach for Large Language Models (LLMs) aimed at aligning more closely with human preferences. By employing the token-wise optimization with forward KL divergence constraints and converting the Bradley-Terry model into a token-level preference model, TDPO addresses key challenges in divergence efficiency and content diversity, surpassing traditional methods like Direct Preference Optimization (DPO) and PPO-based RLHF in tasks such as controlled sentiment generation and single-turn dialogues. This marks a substantial advancement in LLM training methodologies, demonstrating the potential of token-level optimization to enhance the alignment, quality, and diversity of LLM outputs, setting a new direction for AI alignment research and the development of nuanced, human-aligned AI systems.

Regarding the future prospects of alignment methodologies, we anticipate that iterative refinement approaches and multi-turn conversational alignment strategies will significantly improve the alignment of large language models with human values. By continuously refining these models, we can achieve more precise alignment with complex human preferences. Moreover, multi-turn conversations enable deeper and more nuanced interactions, fostering comprehensive attunement to human intentions. These approaches aim to enhance the quality and relevance of AI responses, making AI systems more harmonized with human values and expectations.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D.,

Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code, 2021.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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.

Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked fine-tuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. Pal: Program-aided language models, 2023.

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Havrilla, A., Zhuravinskyi, M., Phung, D., Tiwari, A., Tow, J., Biderman, S., Anthony, Q., and Castricato, L. trlx: A framework for large scale reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8578–8595, 2023.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles,

taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023.

Khalifa, M., Elsahar, H., and Dymetman, M. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.

Knox, W. B., Hatgis-Kessell, S., Booth, S., Niekum, S., Stone, P., and Allievi, A. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*, 2022.

Knox, W. B., Hatgis-Kessell, S., Adalgeirsson, S. O., Booth, S., Dragan, A., Stone, P., and Niekum, S. Learning optimal advantage from preferences and mistaking it for reward. *arXiv preprint arXiv:2310.02456*, 2023.

Koh, H. Y., Ju, J., Liu, M., and Pan, S. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55(8):1–35, December 2022. ISSN 1557-7341. doi: 10.1145/3545176. URL http://dx.doi.org/10.1145/3545176.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

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

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models, 2022. *URL https://arxiv. org/abs/2202.03286*, 2022.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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.

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

Rafailov, R., Hejna, J., Park, R., and Finn, C. From r to Q^*: Your Language Model is Secretly a Q-Function. *arXiv preprint arXiv:2404.12358*, 2024.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Rauh, M., Mellor, J., Uesato, J., Huang, P.-S., Welbl, J., Weidinger, L., Dathathri, S., Glaese, A., Irving, G., Gabriel, I., Isaac, W., and Hendricks, L. A. Characteristics of harmful text: Towards rigorous benchmarking of language models, 2022.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*, 2021.

Song, F., Yu, B., Li, M., Yu, H., Huang, F., Li, Y., and Wang, H. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*, 2023.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback, 2022.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Vu, T.-T., He, X., Haffari, G., and Shareghi, E. Koala: An index for quantifying overlaps with pre-training corpora, 2023.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240*, 2023.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Wiher, G., Meister, C., and Cotterell, R. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022.

Workshop, B., Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Wu, Z., Hu, Y., Shi, W., Dziri, N., Suhr, A., Ammanabrolu, P., Smith, N. A., Ostendorf, M., and Hajishirzi, H. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*, 2023.

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.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zhong, H., Feng, G., Xiong, W., Zhao, L., He, D., Bian, J., and Wang, L. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.

A. Mathematical Derivations

A.1. Proving the Relationship between Maximizing the Advantage Function and Enhancing the Expected Returns

Lemma A.1. *Given two policies π and $\tilde{\pi}$, if for any state $s_t = [x, y^{<t}]$, $\mathbb{E}_{z \sim \tilde{\pi}} \left[A_\pi([x, y^{<t}], z) \right] \geq 0$, then we can conclude:*

$$\mathbb{E}_{x \sim \mathcal{D}} \left[V_{\tilde{\pi}}([x]) \right] \geq \mathbb{E}_{x \sim \mathcal{D}} \left[V_\pi([x]) \right],$$

Proof. Let trajectory $\tau := (x, y^1, y^2, ...)$, and the notation $E_{\tau|\tilde{\pi}}[\cdot]$ indicates that actions are sampled from $\tilde{\pi}$ to generate τ. So we can get

$$\mathbb{E}_{x \sim \mathcal{D}} \left[V_{\tilde{\pi}}([x]) \right] - \mathbb{E}_{x \sim \mathcal{D}} \left[V_\pi([x]) \right] \tag{19}$$

$$= \mathbb{E}_{\tau|\tilde{\pi}} \left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t - V_\pi([x]) \right] \tag{20}$$

$$= \mathbb{E}_{\tau|\tilde{\pi}} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \left(R_t + \gamma V_\pi([x, y^{<t+1}]) - V_\pi([x, y^{<t}]) \right) \right] \tag{21}$$

$$= \mathbb{E}_{\tau|\tilde{\pi}} \left[\sum_{t=1}^{\infty} \gamma^{t-1} A_\pi([x, y^{<t}], y^t) \right] \tag{22}$$

$$= \mathbb{E}_{\tau|\tilde{\pi}} \left[\sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{E}_{y^t \sim \tilde{\pi}} \left[A_\pi([x, y^{<t}], y^t) \right] \right] \tag{23}$$

Since for any state $s_t = [x, y^{<t}]$, $\mathbb{E}_{z \sim \tilde{\pi}} \left[A_\pi([x, y^{<t}], z) \right] \geq 0$, so we can obtain

$$\mathbb{E}_{x \sim \mathcal{D}} \left[V_{\tilde{\pi}}([x]) \right] - \mathbb{E}_{x \sim \mathcal{D}} \left[V_\pi([x]) \right] \geq 0 \tag{24}$$

$\square$

Our goal is to maximize the expected return of a parameterized policy π_θ. According to Eq.23, what we need to do is $\max_{\pi_\theta} \mathbb{E}_{x, y^{<t} \sim \mathcal{D}, z \sim \pi_\theta(\cdot|[x, y^{<t}])} \left[A_{\pi_{\text{ref}}}([x, y^{<t}], z) \right]$. To prevent the excessive degradation of language models, we introduce a reverse KL divergence constraint, forming our objective function:

$$\max_{\pi_\theta} \mathbb{E}_{x, y^{<t} \sim \mathcal{D}, z \sim \pi_\theta(\cdot|[x, y^{<t}])} \left[A_{\pi_{\text{ref}}}([x, y^{<t}], z) - \beta D_{\text{KL}} \left(\pi_\theta(\cdot|[x, y^{<t}]) || \pi_{\text{ref}}(\cdot|[x, y^{<t}]) \right) \right] \tag{25}$$

A.2. Deriving the Mapping between the State-Action Function and the Optimal Policy

Lemma A.2. *The constrained problem in Eq. 7 has the closed-form solution:*

$$\pi_\theta^*(z|[x, y^{<t}]) = \frac{\pi_{\text{ref}}(z|[x, y^{<t}]) \exp \left(\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z) \right)}{Z([x, y^{<t}]; \beta)}, \tag{26}$$

where $Z([x, y^{<t}]; \beta) = \mathbb{E}_{z \sim \pi_{\text{ref}}(\cdot|[x, y^{<t}])} e^{\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z)}$ is the partition function.

Proof.

$$\max_{\pi_\theta} \mathbb{E}_{z \sim \pi_\theta(\cdot | [x, y^{<t}])} A_{\pi_{\text{ref}}}([x, y^{<t}], z) - \beta D_{\text{KL}}\left(\pi_\theta(\cdot | [x, y^{<t}]) \| \pi_{\text{ref}}(\cdot | [x, y^{<t}])\right) \tag{27}$$

$$= \max_{\pi_\theta} \mathbb{E}_{z \sim \pi_\theta(\cdot | [x, y^{<t}])} \left((Q_{\pi_{\text{ref}}}([x, y^{<t}], z) - V_{\pi_{\text{ref}}}([x, y^{<t}])) + \beta \log \left(\frac{\pi_{\text{ref}}(z | [x, y^{<t}])}{\pi_\theta(z | [x, y^{<t}])} \right) \right) \tag{28}$$

$$= \max_{\pi_\theta} \beta \mathbb{E}_{z \sim \pi_\theta(\cdot | [x, y^{<t}])} \log \left(\frac{p(z | [x, y^{<t}]) e^{\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z)}}{\pi_\theta(z | [x, y^{<t}])} \right) - V_{\pi_{\text{ref}}}([x, y^{<t}]) \tag{29}$$

$$= \max_{\pi_\theta} \beta \mathbb{E}_{z \sim \pi_\theta(\cdot | [x, y^{<t}])} \log \left(\frac{\pi_{\text{ref}}(z | [x, y^{<t}]) e^{\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z)}}{Z([x, y^{<t}]; \beta) \pi_\theta(z | [x, y^{<t}])} \right) - V_{\pi_{\text{ref}}}([x, y^{<t}]) + \beta \log Z([x, y^{<t}]; \beta) \tag{30}$$

$$= \max_{\pi_\theta} -\beta D_{\text{KL}}\left(\pi_\theta(z | [x, y^{<t}]) \middle\| \frac{\pi_{\text{ref}}(z | [x, y^{<t}]) e^{\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z)}}{Z([x, y^{<t}]; \beta)} \right) - V_{\pi_{\text{ref}}}([x, y^{<t}]) + \beta \log Z([x, y^{<t}]; \beta) \tag{31}$$

where $Z([x, y^{<t}]; \beta)$ is the partition function:

$$Z([x, y^{<t}]; \beta) = \mathbb{E}_{z \sim \pi_{\text{ref}}(\cdot | [x, y^{<t}])} \exp\left(\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z) \right) \tag{32}$$

Based on the property of KL divergence, we can derive the relationship between the optimal policy and the state-action function:

$$\pi_\theta^*(z | [x, y^{<t}]) = \frac{\pi_{\text{ref}}(z | [x, y^{<t}]) \exp\left(\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z) \right)}{Z([x, y^{<t}]; \beta)} \tag{33}$$

$\square$

A.3. Proving the Equivalence of the Bradley-Terry Model and the Regret Preference Model

Lemma A.3. *Given a reward function $r(\text{x}, \text{y})$, assuming a relationship between token-wise rewards and the reward function represented by $r(x, y) = \sum_{t=1}^{T} \gamma^{t-1} R([x, y^{<t}], y^t)$, we can establish the equivalence between the Bradley-Terry model and the Regret Preference Model in the task of text generation alignment, i.e.,*

$$, \tag{34}$$

where $\sigma(x) = 1/(1 + exp(-x))$ is the logistic sigmoid function.

Proof. According to the Bradley-Terry model, we have

$$P_{\text{BT}}(y_1 \succ y_2 | x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}, \tag{35}$$

where $r(x, y)$ represents the overall reward of the pair (x, y).

Based on assumption that $r(x, y) = \sum_{t=1}^{T} \gamma^{t-1} R([x, y^{<t}], y^t)$, we can get:

$$r(x, y) = \sum_{t=1}^{T} \gamma^{t-1} R([x, y^{<t}], y^t) \tag{36}$$

$$= \sum_{t=1}^{T} \gamma^{t-1} \left(R([x, y^{<t}], y^t) + \gamma V_\pi([x, y^{<t+1}]) - \gamma V_\pi([x, y^{<t+1}]) \right) \tag{37}$$

$$= V_\pi([x, y^{<1}]) + \sum_{t=1}^{T} \gamma^{t-1} \left(R([x, y^{<t}], y^t) + \gamma V_\pi([x, y^{<t+1}]) - V_\pi([x, y^{<t}]) \right) - \gamma^T V_\pi([x, y^{<T+1}]) \tag{38}$$

14

Text generation is analogous to a deterministic contextual bandit, where the transition to the next state is certain given the current state and action, i.e., $p(s_{t+1} = [x, y^{<t+1}]|s_t = [x, y^{<t}], a_t = y^t) = 1$, so we have:

$$Q_\pi([x, y^{<t}], y^t) = R([x, y^{<t}], y^t) + V_\pi([x, y^{<t+1}]) \tag{39}$$

$$A_\pi([x, y^{<t}], y^t) = Q_\pi([x, y^{<t}], y^t) - V_\pi([x, y^{<t}]) \tag{40}$$

Next, note that $y^T = \text{EOS}$ denotes the end of the text sequence. Therefore,

$$V_\pi([x, y^{<T+1}]) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R([x, y^{<T+1+k}], y^{T+1+k}) \middle| s_t = [x, y^{<T+1}] \right] = 0 \tag{41}$$

Substituting Eq.38 to Eq.41 into the Bradley-Terry model, we obtain

$$P_{\text{BT}}(y_1 \succ y_2|x) = \frac{\exp(r(x, y_1))}{\exp(r(x, y_1)) + \exp(r(x, y_2))}$$

$$= \sigma \left(\left(V_\pi([x, y_1^{<1}]) + \sum_{t=1}^{T_1} \left(\gamma^{t-1} A_\pi([x, y_1^{<t}], y^t) \right) \right) - \left(V_\pi([x, y_2^{<1}]) + \sum_{t=1}^{T_2} \left(\gamma^{t-1} A_\pi([x, y_2^{<t}], y_2^t) \right) \right) \right) \tag{42}$$

Additionally, note that $y^{<1} = [\,]$, so we can get

$$V_\pi([x, y_1^{<1}]) = V_\pi([x, [\,]]) = V_\pi([x, y_2^{<1}])$$

Therefore,

$$P_{\text{BT}}(y_1 \succ y_2|x) = \sigma \left(\left(V_\pi([x, y_1^{<1}]) + \sum_{t=1}^{T_1} \left(\gamma^{t-1} A_\pi([x, y_1^{<t}], y_1^t) \right) \right) - \left(V_\pi([x, y_2^{<1}]) + \sum_{t=1}^{T_2} \left(\gamma^{t-1} A_\pi([x, y_2^{<t}], y_2^t) \right) \right) \right) \tag{43}$$

$$= \sigma \left(\sum_{t=1}^{T_1} \left(\gamma^{t-1} A_\pi([x, y_1^{<t}], y_1^t) \right) - \sum_{t=1}^{T_2} \left(\gamma^{t-1} A_\pi([x, y_2^{<t}], y_2^t) \right) \right) \tag{44}$$

$\square$

A.4. Deriving the TDPO Objective Under the Bradley-Terry Model

Theorem A.4. *In the KL-constrained advantage function maximization problem corresponding to Eq.7, the Bradley-Terry model express the human preference probability in terms of the optimal policy π_θ^* and reference policy π_{ref}:*

$$P_{\text{BT}}^*(y_1 \succ y_2|x) = \sigma(u^*(x, y_1, y_2) - \delta^*(x, y_1, y_2)), \tag{45}$$

where, $u(x, y_1, y_2)$ refers to the difference in rewards implicitly defined by the language model π_θ and the reference model π_{ref} (Rafailov et al., 2023), represented as

$$u(x, y_1, y_2) = \beta \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}, \tag{46}$$

and $\delta(x, y_1, y_2)$ refers to the difference in sequential forward KL divergence between two pairs (x, y_1) and (x, y_2), weighted by β, expressed as

$$\delta(x, y_1, y_2) = \beta D_{\text{SeqKL}}(x, y_2; \pi_{\text{ref}} \| \pi_\theta) - \beta D_{\text{SeqKL}}(x, y_1; \pi_{\text{ref}} \| \pi_\theta). \tag{47}$$

Proof. According to the Lemma 4.2, we have

$$\pi_\theta^*(z|[x, y^{<t}]) = \frac{\pi_{\text{ref}}(z|[x, y^{<t}]) \exp\left(\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z) \right)}{Z([x, y^{<t}]; \beta)} \tag{48}$$

where $Z([x, y^{<t}]; \beta) = \mathbb{E}_{z \sim \pi_{\text{ref}}(\cdot|[x, y^{<t}])} e^{\frac{1}{\beta} Q_{\pi_{\text{ref}}}([x, y^{<t}], z)}$ is the partition function.

Rearrange Eq.48, we obtain

$$Q_{\pi_{\text{ref}}}([x, y^{<t}], z) = \beta \log \frac{\pi_\theta^*(z|[x, y^{<t}])}{\pi_{\text{ref}}(z|[x, y^{<t}])} + \beta \log Z([x, y^{<t}]; \beta) \tag{49}$$

From Lemma 4.4, We can get

$$P_{\text{BT}}(y_1 \succ y_2|x) = \sigma \left(\sum_{t=1}^{T_1} \left(\gamma^{t-1} A_\pi([x, y_1^{<t}], y_1^t) \right) - \sum_{t=1}^{T_2} \left(\gamma^{t-1} A_\pi([x, y_2^{<t}], y_2^t) \right) \right) \tag{50}$$

By leveraging Eq.49, we can derive

$$\sum_{t=1}^{T} \gamma^{t-1} A_{\pi_{\text{ref}}}([x, y^{<t}], y^t)$$

$$= \sum_{t=1}^{T} \gamma^{t-1} \left(Q_{\pi_{\text{ref}}}([x, y^{<t}], y^t) - V_{\pi_{\text{ref}}}([x, y^{<t}]) \right) \tag{51}$$

$$= \sum_{t=1}^{T} \gamma^{t-1} \left(Q_{\pi_{\text{ref}}}([x, y^{<t}], y^t) - \mathbb{E}_{z \sim \pi_{\text{ref}}} \left[Q_{\pi_{\text{ref}}}([x, y^{<t}], z) \right] \right) \tag{52}$$

$$= \sum_{t=1}^{T} \gamma^{t-1} \left(\beta \log \frac{\pi_\theta^*(y^t|[x, y^{<t}])}{\pi_{\text{ref}}(y^t|[x, y^{<t}])} + \beta \log Z([x, y^{<t}]; \beta) - \mathbb{E}_{z \sim \pi_{\text{ref}}} \left[\beta \log \frac{\pi_\theta^*(z|[x, y^{<t}])}{\pi_{\text{ref}}(z|[x, y^{<t}])} + \beta \log Z([x, y^{<t}]; \beta) \right] \right) \tag{53}$$

Note that

$$\mathbb{E}_{z \sim \pi_{\text{ref}}} \left[\beta \log Z([x, y^{<t}]; \beta) \right] = \beta \log Z([x, y^{<t}]; \beta) \tag{54}$$

Therefore,

$$\sum_{t=1}^{T} \gamma^{t-1} A_{\pi_{\text{ref}}}([x, y^{<t}], y^t) = \beta \sum_{t=1}^{T} \gamma^{t-1} \left(\log \frac{\pi_\theta^*(y^t|[x, y^{<t}])}{\pi_{\text{ref}}(y^t|[x, y^{<t}])} - \mathbb{E}_{z \sim \pi_{\text{ref}}} \left[\log \frac{\pi_\theta^*(z|[x, y^{<t}])}{\pi_{\text{ref}}(z|[x, y^{<t}])} \right] \right) \tag{55}$$

$$= \beta \sum_{t=1}^{T} \gamma^{t-1} \left(\log \frac{\pi_\theta^*(y^t|[x, y^{<t}])}{\pi_{\text{ref}}(y^t|[x, y^{<t}])} + D_{\text{KL}} \left(\pi_{\text{ref}}(\cdot|[x, y^{<t}]) \| \pi_\theta^*(\cdot|[x, y^{<t}]) \right) \right) \tag{56}$$

$$= \beta \sum_{t=1}^{T} \gamma^{t-1} \log \frac{\pi_\theta^*(y^t|[x, y^{<t}])}{\pi_{\text{ref}}(y^t|[x, y^{<t}])} + \beta \sum_{t=1}^{T} \gamma^{t-1} D_{\text{KL}} \left(\pi_{\text{ref}}(\cdot|[x, y^{<t}]) \| \pi_\theta^*(\cdot|[x, y^{<t}]) \right) \tag{57}$$

When substituting $\gamma = 1$ into the expression, we obtain a more concise form:

$$\sum_{t=1}^{T} A_{\pi_{\text{ref}}}([x, y^{<t}], y^t) = \beta \sum_{t=1}^{T} \log \frac{\pi_\theta^*(y^t|[x, y^{<t}])}{\pi_{\text{ref}}(y^t|[x, y^{<t}])} + \beta \sum_{t=1}^{T} D_{\text{KL}} \left(\pi_{\text{ref}}(\cdot|[x, y^{<t}]) \| \pi_\theta^*(\cdot|[x, y^{<t}]) \right) \tag{58}$$

$$= \beta \left(\log \frac{\pi_\theta^*(y|x)}{\pi_{\text{ref}}(y|x)} + D_{\text{SeqKL}} \left(x, y; \pi_{\text{ref}} \| \pi_\theta^* \right) \right) \tag{59}$$

We let

$$u(x, y_1, y_2) = \beta \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}, \tag{60}$$

$$\delta(x, y_1, y_2) = \beta D_{\text{SeqKL}} \left(x, y_2; \pi_{\text{ref}} \| \pi_\theta \right) - \beta D_{\text{SeqKL}} \left(x, y_1; \pi_{\text{ref}} \| \pi_\theta \right). \tag{61}$$

Substituting Eq.59 to Eq.61 into Eq.50, we arrive at:

$$P_{\text{BT}}^*(y_1 \succ y_2|x) = \sigma(u^*(x, y_1, y_2) - \delta^*(x, y_1, y_2)). \tag{62}$$

$\square$

B. TDPO Implementation Details and Hyperparameters

PyTorch code for the TDPO loss is provided below:

```python
import torch
import torch.nn.functional as F

def tdpo_loss(pi_logits, ref_logits, yw_idxs, yl_idxs, labels, beta, alpha, if_tdpo2):
    """
    pi_logits: policy logits. Shape: (batch_size, sequence_length, vocab_size),
    ref_logits: reference logits. Shape: (batch_size, sequence_length, vocab_size)
    yw_idxs: preferred completion indices in [0,B-1], shape (T,)
    yl_idxs: dispreferred completion indices in [0,B-1], shape (T,)
    labels: labels for which to compute the log probabilities, Shape: (batch_size,
                                        sequence_length)
    beta: temperature controlling strength of KL penalty
    Each pair of (yw_idxs[i], yl_idxs[i]) represents the indices of a single preference
                                        pair.
    alpha: The weight factor adjusts the influence weight of kl divergence at each token
    if_tdpo2: Use method TDPO2 by default; if False, switch to TDPO1
    """

    pi_vocab_logps = pi_logits.log_softmax(-1)

    ref_vocab_ps = ref_logits.softmax(-1)
    ref_vocab_logps = ref_vocab_ps.log()

    pi_per_token_logps = torch.gather(pi_vocab_logps, dim=2, index=labels.unsqueeze(2)).
                                        squeeze(2)
    ref_per_token_logps = torch.gather(ref_vocab_logps, dim=2, index=labels.unsqueeze(2)).
                                        squeeze(2)

    per_position_rewards = pi_per_token_logps - ref_per_token_logps
    yw_rewards, yl_rewards = per_position_rewards[yw_idxs], per_position_rewards[yl_idxs]

    rewards = yw_rewards - yl_rewards

    # losses = -F.logsigmoid(beta * rewards)     # DPO loss function

    # ==============================Difference with DPO==================================
    per_position_kl = (ref_vocab_ps * (ref_vocab_logps - pi_vocab_logps)).sum(-1)
    yw_kl, yl_kl = per_position_kl[yw_idxs], per_position_kl[yl_idxs]

    if not if_tdpo2:
        values = yw_rewards - yl_rewards - (yl_kl - yw_kl)
    else:
        values = yw_rewards - yl_rewards - alpha * (yl_kl - yw_kl.detach())

    losses = -F.logsigmoid(beta * values)
    # ==================================================================================

    return losses
```

Unless specified otherwise, we use a $\alpha = 0.5$, $\beta = 0.1$, batch size of 64, and the RMSprop optimizer with a learning rate of 5e-6. We linearly warm up the learning rate from 0 to 5e-6 over 150 steps.

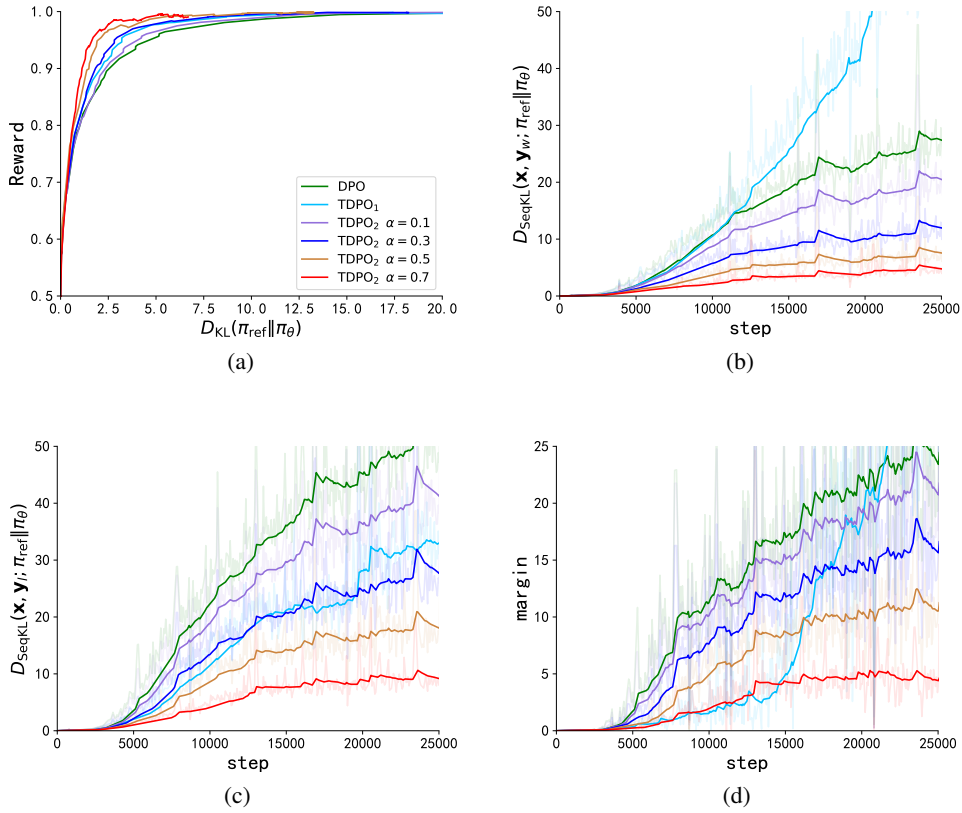# C. Additional Experimental Results

Figure 6. The experiment on IMDb dataset. Figure 6(a) represents the frontier of expected reward and forward KL divergence with respect to the reference model. Figure 6(b) and Figure 6(c) present the progression of sequential forward KL divergence on the preferred and dispreferred responses subset over training steps respectively. Figure 6(d) illustrates the difference between the sequential forward KL divergence on the dispreferred responses subset and that on the preferred responses subset throughout the training process.