Risk-aware Direct Preference Optimization under Nested Risk Measure

Anonymous Authors[1]

Abstract

When fine-tuning pre-trained Large Language Models (LLMs) to align with human values and intentions, the pursuit of maximizing the estimated reward can lead to superior performance, but it also introduces potential risks due to deviations from the original (reference) model's intended behavior. Most existing methods for aligning LLMs typically introduce KL divergence to constrain deviations between the training model and the reference model; however, this may not be sufficient in certain applications that require tight risk control. In this paper, we introduce Risk-aware Direct Preference Optimization (Ra-DPO), a novel approach that incorporates risk-awareness by employing a token-level objective function under nested risk measure. This method formulates a constrained risk-aware advantage function maximization problem and then converts the Bradley-Terry model into a token-level representation. The ultimate objective function maximizes the likelihood of the policy while suppressing the deviation between a training model and the reference model using a sequential risk ratio, thereby enhancing the model's risk-awareness during the process of aligning LLMs. The proposed method's effectiveness is verified via three open-source datasets: IMDb Dataset, Anthropic HH Dataset, and AlpacaEval, and the results demonstrate superior performance of our method in balancing alignment performance and model drift.

1. Introduction

With the advanced and rapid developments of large language models (LLMs) technology, learning from human feedback, serving as a bridge in aligning LLMs with human values and intentions, has become increasingly crucial (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023; Biderman et al., 2023). Reinforcement Learning from Human Feedback (RLHF), which typically involves supervised fine-tuning, reward model training, and further fine-tuning of policy models via reinforcement learning (RL) algorithms, demonstrates impressive capabilities across diverse tasks and has emerged as a concrete research agenda (Christiano et al., 2017; Ouyang et al., 2022; Yuan et al., 2023). A criticized downside is that RLHF has a complex process that requires considerable memory and careful hyperparameter tuning to maintain the stability of RL training.

Direct Preference Optimization (DPO) (Rafailov et al., 2023), featuring a simple and straightforward training process, directly uses the likelihood of the policy to define an implicit reward fitted to the preference data, which has emerged as a popular alternative since it bypasses key challenges in explicit reward modeling and achieves notable efficiency and competitive performance. Nevertheless, some studies (Xiao et al., 2024; Wang et al., 2024b) have reported that DPO still suffers from issues such as excessively long generative responses and the significant KL divergence of the dispreferred response subset. To tackle these issues, numerous variants of DPO have been successively proposed, including f-DPO (Wang et al., 2024a), IPO (Azar et al., 2024), RDPO (Fisch et al., 2024), and SimPO (Meng et al., 2024), which introduce length control mechanisms or enhance KL divergence constraints. However, a key limitation is that these methods only consider evaluation at the sentence level, ignoring the fact that the generation of these responses occurs sequentially, following an auto-regressive approach.

Recently, a fresh perspective on LLMs alignment has been introduced, specifically the sequential and token-level direct preference optimization, known as TDPO (Zeng et al., 2024), which allows for examining divergence in relation to a reference LLM on a more granular, token-by-token basis. Specifically, inspired by Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) in RL field, TDPO redefines the objective of maximizing restricted rewards in a sequential manner and establishes the connection between sentence-level reward and token-level generation through using the Bellman equation. However, since the objective at each step is to maximize the expected return, a risk-neutral criterion, which neglects the characteristics of the reward distribution beyond the mean, TDPO encounters the same

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

challenges as classic RL algorithms (Schulman et al., 2015; 2017; Bisi et al., 2022).

Fortunately, in the field of RL, a series of risk-sensitive methods (Bisi et al., 2022; Candela et al., 2023) have been proposed, which achieve superior performance by introducing various risk measure functions. Recently, some researchers have attempted to introduce this technology in order to align LLMs with human preferences. For instance, RA-RLHF (Chaudhary et al., 2024) introduces Conditional Value at Risk (CVaR) (Artzner, 1997), a static risk measure function, into the fine-tuning of RL, while KTO (Ethayarajh et al., 2024) introduces prospect theory (Tversky & Kahneman, 1992) to fit human choice behavior when faced with uncertain events. However, these methods only analyze the risk of the whole prompt-response at the sentence level by considering the distribution characteristics of the preference data, which neglects the fact that the generation of these responses occurs sequentially, following an auto-regressive approach.

In this paper, we focus on the risk in the value iteration at each step by introducing nested risk measures. Specifically, we investigate a novel direct preference optimization method for the problem of aligning with human preferences from a risk-sensitive perspective and provide corresponding theoretical and empirical results. Our main contributions are summarized as follows.

- We propose a novel Risk-aware Direct Preference Optimization (Ra-DPO) method. This method maximizes the likelihood of the policy while effectively suppressing the deviation between the training model and the reference model by means of a sequential risk ratio, thereby enhancing the model's risk-awareness during the process of balancing alignment performance and model drift.

- We design a new risk-aware token-level objective function by reformulating the constrained reward maximization problem into a token-level form, and then prove that maximizing the objective function will result in policy improvements. Furthermore, by establishing equivalence between the Bradley-Terry model and the Regret Preference Model and deriving the mapping between the risk-aware state-action value function and the optimal policy, we obtain the optimization objective that is solely related to the risk-sensitive policy.

- Experimentally, we provide the results across various text generation tasks to evaluate the effectiveness of our proposed method and the sensitivity to the risk control parameter. The experimental results demonstrate that our method can effectively suppress the risk of model drift while enhancing its performance.

2. Preliminaries

2.1. Preference-based Policy Optimization

Considering a preference-based language model fine-tuning task, let x denote an input prompt (question), and y denote the generated response (answer). The notation $y_w \succ y_l \mid x$ symbolizes the human preference data, where y_w (win) represents a response that is more preferred by humans compared to y_l (lose). Both x and y_w/y_l consist of a sequence of tokens.

Bradley-Terry Model. In the preference-based fine-tuning process, to align with human preferences, a preference predictor adhering to the Bradley-Terry (BT) (Bradley & Terry, 1952) model has been widely employed for pairwise comparisons. The likelihood of a preference pair is commonly expressed using a latent reward model:

$$P_{\mathrm{BT}}\left(y_w \succ y_l \mid x\right)=\frac{\exp\left(r\left(x, y_w\right)\right)}{\exp\left(r\left(x, y_w\right)\right)+\exp\left(r\left(x, y_l\right)\right)}, \tag{1}$$

where $r(x, y_w)$ and $r(x, y_l)$ stand for the reward function at the sentence level from the preferred and dispreferred answers, respectively.

Directly Preference Optimization. Direct Preference Optimization (DPO) (Rafailov et al., 2023) commences with the following RL objective:

$$\begin{aligned}\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} & \left[r\left(x, y\right)\right. \\ & \left.-\beta D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi_{\mathrm{ref}}(\cdot \mid x)\right)\right],\end{aligned} \tag{2}$$

where $\mathcal{D}$ represents the human preference dataset, β is the coefficient of the reverse KL divergence penalty, $\pi_{\mathrm{ref}}\left(\cdot \mid x\right)$ is the policy of fixed reference model (typically selected to be the model that has undergone post-supervised fine-tuning), and $\pi_\theta\left(\cdot \mid x\right)$ represents the policy of the trained model, initialized with $\pi_\theta = \pi_{\mathrm{ref}}$.

By reparameterizing the reward function in Eq. 2 using the policy in a supervised manner, DPO establishes a direct functional mapping between the reward model and the optimal policy.

$$r(x, y)=\beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}+\beta \log Z(x), \tag{3}$$

where $Z(x)$ is the partition function or the normalizing constant.

Then, by plugging the reward from Eq. 3 into the BT model in Eq. 1, DPO derives the objective function:

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

where

$$u\left(x, y_w, y_l\right)=\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)}-\beta \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\mathrm{ref}}\left(y_l \mid x\right)}.$$

2.2. Preference-based Markov Decision Process

A Preference-based Markov Decision Process (Pb-MDP) can be formulated as a modification of the classical MDP: $\mathcal{M}\langle\mathcal{S},\mathcal{A},r,\mathbf{P},\gamma,T\rangle$, where $\mathcal{S}$ and $\mathcal{A}$ represent the finite state and action spaces, respectively; $\mathbf{P}:\mathcal{S}\times\mathcal{A}\to\mathcal{S}$ is the probabilistic transition function; r represents the reward function of the entire prompt-response, which is defined as $(\mathcal{S}\times\mathcal{A})^T\to\mathcal{R}$; γ is the discount factor, and T denotes the length of a trajectory or episode.

Specifically, for language generation, the state $s_t = [x,y^{<t}]\in\mathcal{S}$ is a combination of the prompt and the generated response up to the current step, action $a_t = y^t\in\mathcal{A}$ corresponds to the next generated token, and the token-wise reward is defined as $R_t := R(s_t,a_t) = R([x,y^{<t}],y^t)$. Additionally, note that $y^{<1} = [\ ]$ is an empty sequence. Therefore, we denote $[x]=[x,[\ ]]=\left[x,y^{<1}\right]$. For a given prompt x and the first $t-1$ tokens $y^{<t}$ of the response y, we define the probability distribution of the next token conditioned on $[x,y^{<t}]$ as $\pi_\theta(\cdot\mid[x,y^{<t}])$.

2.3. Risk Measure

It is more desirable to keep risk under control for language generation tasks instead of only considering a risk-neutral criterion, which overlooks the distribution characteristics of rewards, especially on certain safety-critical tasks that may have potential broad societal impact. Therefore, we introduce the risk-sensitive criterion (Bäuerle & Rieder, 2014; Wang & Chapman, 2022) to quantify the hidden risk. More specifically, the definition of the quantile function and risk measure objective are as follows.

The quantile function is the coherent risk-measure (Artzner et al., 1999; Bonetti et al., 2023) of random variable Z,

$$F_Z^{-1}(\xi)=\inf\left\{z\in\mathbb{R}\mid F_Z(z)\geq\xi\right\},$$

which satisfies the following properties for all $Z,Z'\in\mathcal{Z}$: *Concavity:* $\forall\,\lambda\in[0,1]:\eta(\lambda Z+(1-\lambda)Z')\geq\lambda\eta(Z)+(1-\lambda)\eta(Z')$; *Monotonicity:* If $Z\geq Z'$, then $\eta(Z)\geq\eta(Z')$; *Translation Equivariance:* $\forall\,\epsilon\in\mathbb{R}:\eta(Z+\epsilon)=\eta(Z)+\epsilon$; *Positive Homogeneity:* $\forall\,\lambda>0:\eta(\lambda Z)=\lambda\eta(Z)$. Then, we introduce the nested risk-measures that are built upon Pb-MDP in Subsection 2.3.

Nested risk-measures. In the context of standard Pb-MDP, the nested quantile risk measures (Fei et al., 2020; Chen et al., 2024; Zhao et al., 2024) can be elucidated in Bellman equation type as follows:

$$\begin{cases}Q_\pi([x,y^{<t}],y^t)=R([x,y^{<t}],y^t)+\Phi^\mu(V_\pi([x,y^{<t}])),\\ V_\pi([x,y^{<t}])=Q_\pi([x,y^{<t}],\pi(\cdot\mid[x,y^{<t}])),\\ V_\pi\left(\left[x,y^{<T}\right]\right)=R\left(\left[x,y^{<T}\right]\right),\end{cases}\tag{5}$$

where $Q_\pi([x,y^{<t}],y^t)$ and $V_\pi([x,y^{<t}])$ represent the state-action value and state value under the nested risk measures at timestep $t\in[1,\cdots,T]$, respectively. $\Phi(\cdot)$ is a nested risk measure function with a risk control parameter μ. For any random variable Z, we have

$$\Phi^\mu(Z)=\int_0^1 F_Z^{-1}(\xi)\mathrm{d}G(\xi),$$

where G is a weighting function over the quantiles.

This class captures a broad range of useful objectives, including the popular CVaR (Artzner, 1997) objective. Due to space constraints, we provide a detailed survey about risk measure in Appendix A.1 and the expanded version of value function definition in Appendix A.2.

3. Methodology

This section proposes a novel language model alignment method called Risk-aware Direct Preference Optimization (Ra-DPO). Specifically, we first conduct an analysis of the characteristics of nested risk measures and design a new risk-aware token-level objective function by reformulating the constrained reward maximization problem into a token-level form. Subsequently, we prove that maximizing the objective function will result in policy improvements. Then, the optimization objective solely related to the risk-sensitive policy is obtained by deriving the mapping between the risk-aware state-action function and the optimal policy; and establishing BT model equivalence with the Regret Preference Model. Finally, we conduct a formalized analysis of this optimization objective in terms of derivatives and derive the loss function for Ra-DPO.

3.1. Risk-aware Objective Function

In this subsection, we aim to design a new risk-aware objective function for preference-based language model fine-tuning. Unfortunately, although the recursive Bellman equation under nested risk measures was introduced in Subsection 2.3, it cannot be directly applied, mainly due to the following reasons:

(1) For the Pb-MDP setting, the algorithm can only obtain the reward (an implicit reward fitted to the preference data) at an entire prompt-response until the end and thus cannot compute the target value at each step.

(2) The nested risk-measures incorporate a Bellman-type recursion and are not law-invariant (Hau et al., 2023), which are complex and difficult to compute.

To surmount these obstacles, a straightforward approach is to introduce the state augmentation method, i.e., reconstructing an augmented Pb-MDP as described in (Zhao et al., 2024), where the state at each timestep includes historical trajectories. This method can reformulate the recursive

Bellman equation into a classical Bellman equation with augmented states. However, it is noteworthy that, in this paper, we directly define the state as a combination of the prompt and the generated response up to the current step to model the sequential and auto-regressive generation. It possesses a characteristic in that the state at the previous timestep is a subset of the state at the current timestep, i.e., $\left[x, y^{<t-1}\right] \subset [x, y^{<t}]$. Therefore, we can rewrite the nested quantile objective's Bellman equation in Eq. 5 as follows:

$$\begin{cases} \tilde{Q}_\pi\left([x, y^{<t}], y^t\right) = \Phi^\mu\left(\tilde{V}_\pi\left(y^{t+1} \circ \left([x, y^{<t}], y^t\right)\right)\right), \\ \tilde{V}_\pi\left([x, y^{<t}]\right) = \tilde{Q}_\pi\left([x, y^{<t}], \pi\left(\cdot \mid [x, y^{<t}]\right)\right), \\ \tilde{V}_\pi\left(\left[x, y^{<T}\right]\right) = R\left(\left[x, y^{<T}\right]\right), \end{cases} \tag{6}$$

where $\tilde{Q}_\pi\left([x, y^{<t}], y^t\right)$ and $\tilde{V}_\pi\left([x, y^{<t}]\right)$ represent the risk-aware state value and state-action value under the policy π, respectively.

It is noteworthy that there is a significant difference in the calculation of the risk-aware state value function between Eq. 5 and Eq. 6. And, according to the Lemma 3.6 in (Zhao et al., 2024), we can obtain the following lemma.

Lemma 3.1. *For a given Pb-MDP, the reward on the entire prompt-response can be decomposed as $r = \sum_{t=1}^{T} \gamma^{t-1} R\left([x, y^{<t}], y^t\right)$, the relationship between the state value function Eq. 5 and Eq. 6 is as follows:*

$$\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right) = V_\pi\left(\left[x, y^{<t}\right]\right) + R_{1:t-1}, \tag{7}$$

where $R_{1:t-1} = \sum_{h=1}^{t-1} \gamma^{h-1} R\left(\left[x, y^{<h}\right], y^h\right)$ denotes the reward of the $1 \sim t-1$ steps of a prompt-response, and $V_\pi[x]$ and $\tilde{V}_\pi[x]$ are equivalent.

The proof is detailed in Appendix B.1.

Subsequently, based on the new risk-aware state value and state-action value in Eq. 6, we define the risk-aware advantage function as follows.

Definition 3.2. For a risk-sensitive Pb-MDP that satisfies the Bellman equation in Eq. 6, the risk-aware advantage function can be defined as

$$\tilde{A}_\pi\left(\left[x, y^{<t}\right], z\right) = \tilde{Q}_\pi\left(\left[x, y^{<t}\right], z\right) - \Phi^\mu(\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)), \tag{8}$$

where z subject to $\pi_\theta\left(\cdot \mid [x, y^{<t}]\right)$.

The definition is reasonable, and the derivation provided in Appendix B.2.

Furthermore, based on the definition of risk-aware advantage function in Definition 3.2, we propose a new risk-aware objective function:

$$\begin{aligned} \max_{\pi_\theta} \mathbb{E}_{x, y^{<t} \sim \mathcal{D}, z \sim \pi_\theta(\cdot \mid [x, y^{<t}])} \Big[& \tilde{A}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right) \\ & - \beta D_{\text{KL}}\left(\pi_\theta\left(\cdot \mid \left[x, y^{<t}\right]\right) \| \pi_{\text{ref}}\left(\cdot \mid \left[x, y^{<t}\right]\right)\right) \Big]. \end{aligned} \tag{9}$$

The objective function maximizes a risk-sensitive advantage function subject to a KL divergence constraint, which takes into account the risk when selecting the optimal policy, thereby achieving a better balance between alignment performance and model drift. Next, we prove that maximizing the risk-aware objective function in Eq. 9 will result in policy improvements, as stated in the following lemma.

Lemma 3.3. *Given two policies π and π', if for any state $s_t = [x, y^{<t}]$, $\mathbb{E}_{z \sim \pi'}\left[\tilde{A}_\pi\left([x, y^{<t}], z\right)\right] \geq 0$, then we can conclude:*

$$\mathbb{E}_{x \sim \mathcal{D}}\left[\tilde{V}_{\pi'}([x])\right] \geq \mathbb{E}_{x \sim \mathcal{D}}\left[\tilde{V}_\pi([x])\right]. \tag{10}$$

The proof is provided in Appendix B.3.

3.2. Risk-aware Preference Optimization

In this subsection, we focus on how to convert the BT model into risk-sensitive token-level representation to obtain the optimization objective that is solely related to the risk-sensitive policy, which is divided into two steps: (1) derive the mapping between the risk-aware state-action function and the optimal policy; (2) establish BT model equivalence with the Regret Preference Model.

Specifically, starting from the risk-aware token-level objective function in Eq. 9, we first derive the mapping between the risk-aware state-action function $\tilde{Q}_\pi$ and the optimal policy π_θ^*, as stated in the following lemma.

Lemma 3.4. *The constrained problem in Eq. 9 has the closed-form solution:*

$$\begin{aligned} & \pi_\theta^*\left(z \mid \left[x, y^{<t}\right]\right) \\ & \quad = \frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right) \exp\left(\frac{1}{\beta} \tilde{Q}_{\pi_{\text{ref}}}\left([x, y^{<t}], z\right)\right)}{Z\left([x, y^{<t}]; \beta\right)}, \end{aligned} \tag{11}$$

where

$$Z\left(\left[x, y^{<t}\right]; \beta\right) = \mathbb{E}_{z \sim \pi_{\text{ref}}(\cdot \mid [x, y^{<t}])} e^{\frac{1}{\beta} \tilde{Q}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right)},$$

which is the partition function.

The proof is provided in Appendix B.4. Then, by rearranging Eq. 11, we can obtain the expression of the risk-aware state-action function in terms of the policy

$$\begin{aligned} & \tilde{Q}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right) \\ & \quad = \beta \log \frac{\pi_\theta^*\left(z \mid [x, y^{<t}]\right)}{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right)} + \beta \log Z\left(\left[x, y^{<t}\right]; \beta\right). \end{aligned} \tag{12}$$

Subsequently, by utilizing the reward decomposition formula $r = \sum_{t=1}^{T} \gamma^{t-1} R\left([x, y^{<t}], y^t\right)$ from Lemma 3.1, we establish BT model equivalence with the Regret Preference Model, as shown in the following lemma.

Lemma 3.5. *Given a reward function r of the entire prompt-response, based on a relationship between token-wise rewards and the reward function represented by $r = \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right)$, we can establish the equivalence between the Bradley-Terry model and the Regret Preference Model, i.e.,*

$$P_{\mathrm{BT}}\left(y_1 \succ y_2 \mid x\right) = \sigma\left(\sum_{t=1}^{T_1} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_1^{<t}\right], y_1^t\right) - \sum_{t=1}^{T_2} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_2^{<t}\right], y_2^t\right)\right), \tag{13}$$

where $\sigma(z) = 1/\left(1+\exp(-z)\right)$ is the logistic sigmoid function for any random variable z.

The proof is provided in Appendix B.5.

According to the definition of the risk-aware advantage function in Definition 3.2, we can directly establish the relationship between the optimal solution in Eq. 12 and preference optimization objective in Eq. 13. In this way, we ultimately reformulate the BT model to be directly tied to the risk-aware optimal policy π_θ^* and the reference policy π_{ref}, which is summarized in the following theorem.

Theorem 3.6. *Given prompts x and pairwise responses (y_1, y_2), and the risk-aware objective function in Eq. 9, the Bradley-Terry model expresses the human preference probability in terms of the risk-aware optimal policy π_θ^* and reference policy π_{ref}:*

$$P_{\mathrm{BT}}^*\left(y_1 \succ y_2 \mid x\right) = \sigma\left(u^*\left(x, y_1, y_2\right) - \delta^*\left(x, y_1, y_2\right)\right), \tag{14}$$

where $u\left(x, y_1, y_2\right)$ represents the difference in implicit rewards defined by the risk-aware policy π_θ^* and the reference policy π_{ref}, weighted by β, represented as

$$u\left(x, y_1, y_2\right) = \beta \log \frac{\pi_\theta\left(y_1 \mid x\right)}{\pi_{\mathrm{ref}}\left(y_1 \mid x\right)} - \beta \log \frac{\pi_\theta\left(y_2 \mid x\right)}{\pi_{\mathrm{ref}}\left(y_2 \mid x\right)}, \tag{15}$$

and $\delta\left(x, y_1, y_2\right)$ represents the difference in sequential risk ratio between two pairs (x, y_1) and (x, y_2), expressed as

$$\begin{aligned}\delta\left(x, y_1, y_2\right) = & \beta D_{\mathrm{SeqRR}}\left(x, y_2; \pi_{\mathrm{ref}} \mid \pi_\theta\right) \\ & - \beta D_{\mathrm{SeqRR}}\left(x, y_1; \pi_{\mathrm{ref}} \mid \pi_\theta\right),\end{aligned} \tag{16}$$

where

$$D_{\mathrm{SeqRR}}\left(x, y; \pi_{\mathrm{ref}} \mid \pi_\theta\right) = \sum_{t=1}^{T} \Phi^{\mu}_{z \sim \pi_{\mathrm{ref}}}\left(\log \frac{\pi_{\mathrm{ref}}\left(z \mid x\right)}{\pi_\theta\left(z \mid x\right)}\right).$$

The proof is provided in the Appendix B.6.

3.3. Loss Function and Formal Analysis

Drawing on Theorem 3.6, we reformulate the BT model into a structure solely relevant to the risk-sensitive policy, which enables us to formulate a likelihood maximization objective for a parametrized policy π_θ, and then our loss function becomes:

$$\begin{aligned}&\mathcal{L}_{\text{Ra-DPO}_1}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) \\ &\quad = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma\left(u\left(x, y_w, y_l\right) - \delta\left(x, y_w, y_l\right)\right)\right].\end{aligned} \tag{17}$$

Through this approach, we explicitly introduce sequential risk ratio into the loss function, which incorporates risk-awareness during the process of balancing alignment performance and model drift. To elucidate the benefit of the proposed method, we give further interpretation by analyzing the loss function and its gradient. Specifically, we conduct a derivative analysis of our method. For convenience, we use u to denote $u\left(x, y_w, y_l\right)$, and δ to represent $\delta\left(x, y_w, y_l\right)$. By simple calculations, we can derive the gradient of the loss function in Eq. 17 with respect to the parameters θ :

$$\begin{aligned}&\nabla_\theta \mathcal{L}_{\text{Ra-DPO}_1}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) \\ &\quad = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\left(-u + \delta\right)\left[\nabla_\theta u - \nabla_\theta \delta\right]\right],\end{aligned} \tag{18}$$

where $(-u + \delta)$ serves as the weighting factor for the gradient.

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

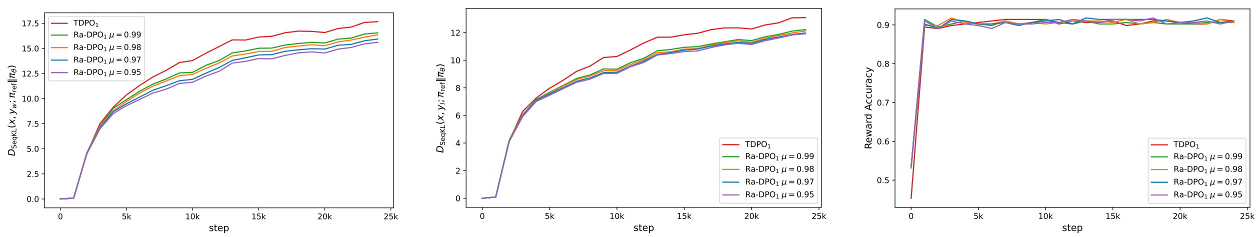
$$\begin{aligned}\mathcal{L}_{\mathrm{TDPO}_2}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) = -\mathbb{E}\Bigg[\log \sigma\Bigg(\left(\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)} - \beta \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\mathrm{ref}}\left(y_l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

$$\begin{aligned}\mathcal{L}_{\text{Ra-DPO}_2}\left(\pi_\theta; \pi_{\mathrm{ref}}\right) = -\mathbb{E}\Bigg[\log \sigma\Bigg(\left(\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\mathrm{ref}}\left(y_w \mid x\right)} - \beta \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\mathrm{ref}}\left(y_l \mid x\right)}\right) \\ -\alpha\left(\beta D_{\mathrm{SeqRR}}\left(x, y_l; \pi_{\mathrm{ref}} \mid \pi_\theta\right) - \mathrm{sg}\left(\beta D_{\mathrm{SeqRR}}\left(x, y_w; \pi_{\mathrm{ref}} \mid \pi_\theta\right)\right)\right)\Bigg)\Bigg]\end{aligned}$$

Figure 1. Comparison of loss functions for DPO, TDPO_2 and Ra-DPO_2 methods. The sg denotes the stop-gradient operator.

From Eq. 18, we can observe that the first part $(-u)$ corresponds to the weight factor in the first part of loss function of TDPO. Its value will increase when the language model makes prediction errors relative to human preferences, i.e., $\log \frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)} > \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}$. The second part δ consists of the difference between the sequential risk ratio of the dispreferred response subset and the preferred response subset, which is a distinctive component of our method. When selecting a convex function (risk-averse), such as CVaR, as the risk measure function, our method automatically balances the risk ratio.

Furthermore, based on a common starting point shared by our method and TDPO (Zeng et al., 2024), i.e., reducing risks stemming from model drift and ensuring training stability, we also provide the second version of our method,



(a) (b) (c)

Figure 2. The experiment on the IMDb dataset with GPT-2 Large serving as the base model. Figure 2(a) and Figure 2(b) present the progression of sequential KL divergence (the lower the better) of both preferred response and dispreferred responses. Additionally, Figure 2(c) illustrates the reward accuracy curves (the higher the better).

Ra-DPO$_2$. The loss function of Ra-DPO$_2$ is given by:

$$\mathcal{L}_{\text{Ra-DPO}_2}(\pi_\theta;\pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(u(x,y_w,y_l)-\alpha\delta_2(x,y_w,y_l)\right)\right], \tag{19}$$

where α is a parameter, and

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

The sg represents the stop-gradient operator, which blocks the propagation of gradients. Ra-DPO$_2$ modifies the loss function of Ra-DPO$_1$ by discontinuing the gradient propagation of $D_{\text{SeqRR}}(x,y_w;\pi_{\text{ref}}\mid\pi_\theta)$ and treating it as a baseline term for alignment of $D_{\text{SeqRR}}(x,y_l;\pi_{\text{ref}}\mid\pi_\theta)$. The aim of the modification is to ensure training stability, rather than accelerating the training speed.

To summarize, the comparison of the loss functions for DPO, TDPO$_2$, and Ra-DPO$_2$ is shown in Figure 1. In addition, we give a procedure of our method, and provide its pseudocode (Algorithm 1) in Appendix B.7.

4. Experiments

We empirically evaluate our method via several open-source datasets and pre-trained models. Our experiments aim to answer the following questions: First, how does the performance of our method compare with existing methods, and is our method more sensitive to risks when tackling challenging text generation tasks? Second, how does the risk control parameter μ affect the performance of our method?

To answer these questions, we conduct experiments on IMDb Dataset (Maas et al., 2011), Anthropic HH Dataset (Bai et al., 2022) and AlpacaEval (Dubois et al., 2024) for three different text generation tasks. Based on the original *KTO implementation*[1], we trained Ra-DPO and the baseline models using the same hyperparameters. Specifically, for Ra-DPO, we employed the popular CVaR (Artzner, 1997) as the risk measure function. We compare our method against the following algorithms: (1) DPO (Rafailov et al., 2023), which only considers evaluation at the sentence level; (2) PPO (Schulman et al., 2017), which is an offline PPO variant provided by the original KTO implementation; (3) TDPO$_1$ and TDPO$_2$ (Zeng et al., 2024), which convert the BT model into token-level representation to obtain the optimization objective; (4) KTO (Ethayarajh et al., 2024), which considers humans make decisions that do not maximize their expected value when faced with uncertain events. All reported results of our algorithm and baseline algorithms are trained using 4 $\times$ A100 GPUs, each with 40GB of memory.

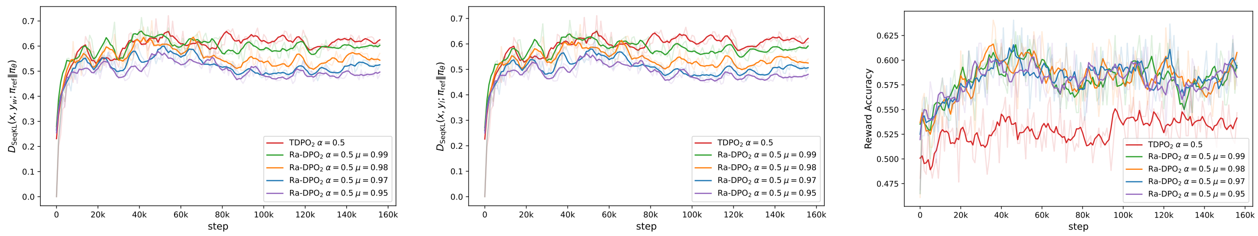
4.1. Experiments on IMDb Dataset

Experimental setup: The IMDb dataset is a controlled semantic generation dataset within the context of movie reviews, serving as a valuable resource for training and evaluating sentiment analysis models. We employ GPT-2 Large (Radford et al., 2019) as our base model and use the model checkpoint: *insub/gpt2-large-IMDb-fine-tuned*[2] as the SFT model. In this setup, the model is presented with prompts consisting of prefixes from movie reviews, and is required to generate responses with positive sentiment. Specifically, we implement the versions of Ra-DPO$_1$ with risk control parameter $\mu\in\{0.99,0.98,0.97,0.95\}$. Moreover, in order to achieve a fair comparison, we calculate the sequential KL divergence for our method. Note that the risk ratio value is slightly larger than the KL divergence value when selecting CVaR (a convex function) as the risk measure function. The results are shown in Figure 2.

Evaluation: Figure 2 shows that Ra-DPO$_1$ can outperform or achieve reward accuracy similar to the advanced TDPO algorithm while also maintaining a slight model drift (indicated by the lower KL divergence), demonstrating the

[1]Available at `https://github.com/ContextualAI/HALOs`

[2]https://huggingface.co/insub/gpt2-large-IMDb-fine-tuned



(a) (b) (c)

Figure 3. The experiment on the Anthropic HH dataset with Pythia-1.4B serving as the base model. We implemented TDPO$_2$, and different versions of Ra-DPO$_2$ with respect to the risk control parameter μ while keeping coefficient α constant at 0.5. Figure 3(a) and Figure 3(b) present the progression of sequential KL divergence (the lower the better) of both preferred response and dispreferred responses. Additionally, Figure 3(c) illustrates the reward accuracy curves (the higher the better).

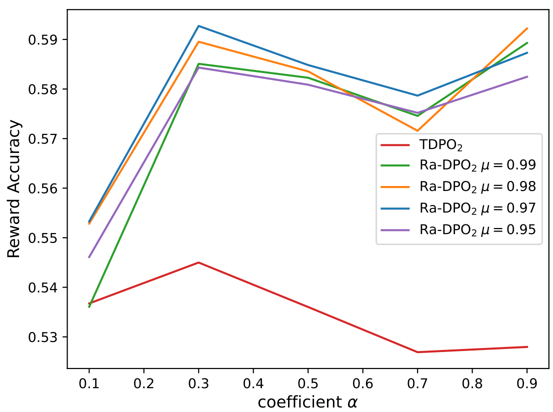


Figure 4. The reward accuracy of each algorithm on the Anthropic HH dataset, using Pythia-1.4B as the base model.

risk-awareness of Ra-DPO$_1$ during the process of balancing alignment performance and model drift.

4.2. Experiments on Anthropic HH Dataset

Experimental setup: Anthropic HH dataset contains 170k dialogues between a human and an automated assistant, where each transcript ends with a pair of responses generated by an LLM along with a preference label denoting the human-preferred response. We use Pythia-1.4B and Pythia-2.8B (Biderman et al., 2023) as the base models to test our method on Anthropic HH dataset, respectively. Here, the reference models are trained by fine-tuning the base models on chosen completions. Specifically, we implement TDPO$_2$ and different versions of Ra-DPO$_2$ with respect to the parameters μ and α The results are depicted in Figure 3, Figure 4, and Appendix C.1.

Evaluation: Figure 3 shows the performance of TDPO$_2$, and different versions of Ra-DPO$_2$ with respect to the risk control parameter μ while keeping coefficient α constant at

Table 1. AlpacaEval compares the responses generated by Algorithms DPO, PPO, KTO, TDPO$_1$, TDPO$_2$ ($\alpha = 0.5$), Ra-DPO$_1$ ($\mu = 0.97$), and Ra-DPO$_2$ ($\alpha = 0.5$, $\mu = 0.97$) with those generated by *gpt4_1106_preview*. The winrate and length-controlled winrate (Lc winrate) are evaluated based on *oasst_pythia_12b*.

Method	Winrate	Lc winrate
DPO	51.1± 1.9	44.7± 0.4
PPO	52.1± 1.8	51.9± 0.5
KTO	51.5± 1.8	50.2± 0.6
TDPO$_1$	51.9± 1.8	53.0± 0.6
TDPO$_2$	52.2± 1.6	52.2± 0.5
Ra-DPO$_1$	**53.5± 1.8**	53.9± 0.5
Ra-DPO$_2$	52.1± 1.8	**55.7± 0.5**

0.5. From the figure, we notice that Ra-DPO$_2$ achieves superior performance (the higher reward accuracy) and maintains a slight model drift (the lower KL divergence). Figure 4 shows the reward accuracy of responses generated by models trained with different algorithms. The results demonstrate that when the coefficient $\alpha > 0.1$, the reward accuracy of Ra-DPO$_2$ exceeds that of TDPO$_2$ across all risk control parameter μ. These results demonstrate that Ra-DPO$_2$ possesses a strong capability to align with human preferences.

4.3. Experiments on AlpacaEval

Experimental setup: To comprehensively evaluate the performance of Ra-DPO$_2$, we conducted pairwise comparisons on AlpacaEval using models trained on Anthropic HH dataset. Following the official *AlpacaEval implementation*[3], we sampled responses with a temperature coefficient of 0.7. The comparisons about winrate based on *oasst_pythia_12b*[4] are summarized in Table 1 and Figure 5.

[3]https://github.com/tatsu-lab/alpaca_eval

[4]https://huggingface.co/OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5

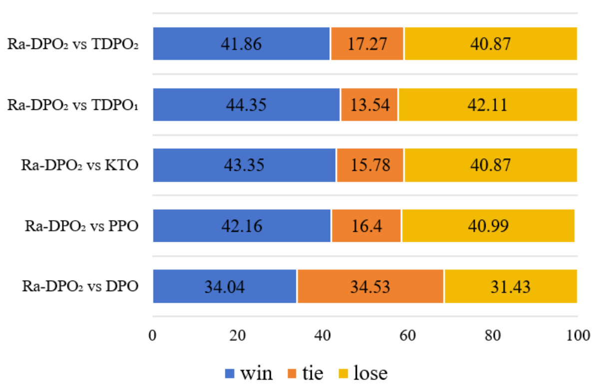


Figure 5. AlpacaEval comparison between DPO, PPO, TDPO_1, TDPO_2 and Ra-DPO_2 methods. The win, tie, and lose rates are evaluated based on *oasst-pythia-12b*.

Evaluation: Table 1 reveals that under the two indicators of winrate and length-controlled winrate, most of the implemented algorithms can outperform the common default baseline *gpt4_1106_preview* (DPO is more prone to generating long responses). Among them, Ra-DPO_1 and Ra-DPO_2 demonstrate the highest level of performance, especially when it comes to the length-controlled winrate indicator. Figure 5 presents a straightforward result: Compared to the baseline algorithms, Ra-DPO_2 achieves a high winrate, demonstrating superior performance in assisting LLMs to generate high-quality responses.

5. Related Work

5.1. LLMs Alignment

During the development and implementation of LLMs, numerous researchers have encountered challenges in balancing adherence to human instructions (explicit objective) with the pursuit of being helpful, honest, and harmless (implicit objectives), challenges that stem from the misaligned next token prediction task used in the pre-training stage (Bai et al., 2022; Bhardwaj & Poria, 2023; Dai et al., 2024; Yeh et al., 2024). Therefore, a typical post-training stage, referred to as preference optimization (e.g., RLHF and DPO), is additionally performed to align pre-trained language models with human intentions, and it has become a crucial aspect in the fine-tuning of large models, often indispensable. Currently, most approaches (Wu et al., 2023; Wang et al., 2024a; Meng et al., 2024) utilize KL divergence at the sentence level to ensure that the training model remains closely aligned with a reference model, preventing significant deviations. However, the generation of these responses occurs sequentially, following an auto-regressive approach. Recent works (Zeng et al., 2024; Ouyang et al., 2024) introduce a fresh perspective, specifically the sequential and token-level direct preference optimization, which allows for examining KL divergence in relation to a reference LLM on a more granular, token-by-token basis. However, due to the neglect of the characteristics of a reward distribution other than the mean, these methods still suffer from the trouble of being insensitive to risk.

5.2. Risk-aware Reinforcement Learning

Reinforcement learning has made groundbreaking achievements through approaches such as Q-learning (Mnih et al., 2015) and policy gradients (Schulman et al., 2015; 2017) in sequential decision tasks, but it also faces challenges when considering application in the real world (Mnih et al., 2015; Wang & Chapman, 2022). A primary reason is that the risk-neutral criterion (maximizing the expectation) ignores the characteristics of a reward distribution other than the mean, which may be important for systems with safety concerns, especially in certain applications requiring tight risk control (Fei et al., 2020; Bisi et al., 2022). In order to tackle this challenge, two types of risk-sensitive measures have been introduced: nested and static quantile risk-aware measures. Static risk measures (Fei et al., 2021; Wang et al., 2023) are straightforward to interpret, but the resulting optimal policy may not remain Markovian and may become history-dependent. On the other hand, nested risk measures (Chen et al., 2024; Zhao et al., 2024) utilize MDPs to ensure risk sensitivity of the value iteration at each step under the current state, resulting in a more conservative approach. In this paper, we prefer nested risk measures because they recursively adhere to the Bellman equation and allow the MDPs to be reconstructed through state augmentation, enabling them to remain Markovian and ensuring that policy choices depend solely on the current state.

6. Conclusion

A pressing challenge arises for language generation tasks in the area of risk control, as the models, once trained, are often required to interact directly with humans. In this paper, we propose a novel direct preference optimization method that incorporates risk awareness by introducing nested risk measures into the Bellman equation, to align pre-trained LLMs with human preferences. Specifically, we design a new risk-aware token-level objective function by reformulating the constrained reward maximization problem into a token-level form and then prove that maximizing this objective function leads to improvements in policy performance. Then, an optimization objective solely related to the risk-sensitive policy is obtained by deriving the mapping between the risk-aware state-action function and the optimal policy and establishing BT model equivalence with the Regret Preference Model. Finally, we conduct a formal analysis of this optimization objective and derive the loss function of Ra-DPO, which has practical implications for language generation tasks.

Impact Statement

This paper presents work that aims to make LLMs more helpful and safer. Our work has many positive societal impacts, such as providing a theoretical foundation for risk-aware language generation task, none of which we feel must be specifically highlighted. There are no negative societal impacts on our work.

References

Artzner, P. Thinking coherently. *Risk*, 10:68–71, 1997.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, 2024.

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.

Bäuerle, N. and Rieder, U. More risk-sensitive markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

Bhardwaj, R. and Poria, S. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*, 2023.

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 *ICML*, 2023.

Bisi, L., Santambrogio, D., Sandrelli, F., Tirinzoni, A., Ziebart, B. D., and Restelli, M. Risk-averse policy optimization via risk-neutral policy optimization. *Artificial Intelligence*, 311:103765, 2022.

Bonetti, M., Bisi, L., and Restelli, M. Risk-averse optimization of reward-based coherent risk measures. *Artificial Intelligence*, 316:103845, 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.

Candela, E., Doustaly, O., Parada, L., Feng, F., Demiris, Y., and Angeloudis, P. Risk-aware controller for autonomous vehicles using model-based collision prediction and reinforcement learning. *Artificial Intelligence*, 320:103923, 2023.

Chaudhary, S., Dinesha, U., Kalathil, D., and Shakkottai, S. Risk-averse fine-tuning of large language models. In *NeurIPS*, 2024.

Chen, Y., Du, Y., Hu, P., Wang, S., Wu, D., and Huang, L. Provably efficient iterated cvar reinforcement learning with function approximation and human feedback. In *ICLR*, 2024.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *NeurIPS*, 2017.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe rlhf: Safe reinforcement learning from human feedback. In *ICLR*, 2024.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Model alignment as prospect theoretic optimization. In *ICML*, 2024.

Fei, Y., Yang, Z., Chen, Y., Wang, Z., and Xie, Q. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *NeurIPS*, 2020.

Fei, Y., Yang, Z., and Wang, Z. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *ICML*, 2021.

Fisch, A., Eisenstein, J., Zayats, V., Agarwal, A., Beirami, A., Nagpal, C., Shaw, P., and Berant, J. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in markov decision processes. *Artificial intelligence*, 147(1-2):163–223, 2003.

Hau, J. L., Petrik, M., and Ghavamzadeh, M. Entropic risk optimization in discounted mdps. In *AISTATS*, pp. 47–76, 2023.

Huber, J., Payne, J. W., and Puto, C. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research*, 9 (1):90–98, 1982.

Lowd, D. and Davis, J. Learning markov network structure with decision trees. In *ICDM*, 2010.

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *ACL*, pp. 142–150, 2011.

Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Ouyang, Y., Wang, L., Yang, F., Zhao, P., Huang, C., Liu, J., Pang, B., Yang, Y., Zhan, Y., Sun, H., et al. Token-level proximal policy optimization for query generation. *arXiv preprint arXiv:2411.00722*, 2024.

Peuter, S. D., Zhu, S., Guo, Y., Howes, A., and Kaski, S. Preference learning of latent decision utilities with a human-like model of preferential choice. In *NeurIPS*, 2024.

Pichler, A. and Schlotter, R. Entropy based risk measures. *European Journal of Operational Research*, 285(1):223–236, 2020.

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. In *NeurIPS*, 2023.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *ICML*, 2015.

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

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.

Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5:297–323, 1992.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *ICLR*, 2024a.

Wang, K., Kallus, N., and Sun, W. Near-minimax-optimal risk-sensitive reinforcement learning with cvar. In *ICML*, 2023.

Wang, Y. and Chapman, M. P. Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control. *Artificial Intelligence*, 311:103743, 2022.

Wang, Z., Bi, B., Pentyala, S. K., Ramnath, K., Chaudhuri, S., Mehrotra, S., Mao, X.-B., Asur, S., et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024b.

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. In *NeurIPS*, 2023.

Xiao, W., Wang, Z., Gan, L., Zhao, S., He, W., Tuan, L. A., Chen, L., Jiang, H., Zhao, Z., and Wu, F. A comprehensive survey of datasets, theories, variants, and applications in direct preference optimization. *arXiv preprint arXiv:2410.15595*, 2024.

Yeh, M.-H., Tao, L., Wang, J., Du, X., and Li, Y. How reliable is human feedback for aligning large language models? *arXiv preprint arXiv:2410.01957*, 2024.

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.

Zeng, Y., Liu, G., Ma, W., Yang, N., Zhang, H., and Wang, J. Token-level direct preference optimization. In *ICML*, 2024.

Zhang, L., Li, L., Wei, W., Song, H., Yang, Y., and Liang, J. Scalable constrained policy optimization for safe multi-agent reinforcement learning. In *NeurIPS*, 2024.

Zhao, W., He, T., and Liu, C. Model-free safe control for zero-violation reinforcement learning. In *CoRL*, 2021.

Zhao, Y., Escamilla, J. E. A., Lu, W., and Wang, H. Ra-pbrl: Provably efficient risk-aware preference-based reinforcement learning. In *NeurIPS*, 2024.

A. Supplementary Materials for Section 2

A.1. Risk Measure: A Brief Overview

Risk-aware Reinforcement Learning. Reinforcement learning has made groundbreaking achievements through approaches such as Q-learning (Mnih et al., 2015) and policy gradients (Schulman et al., 2015; 2017) in sequence decision tasks and has been gradually maturing in laboratory-level applications. In recent years, many researchers have gradually shifted their attention to real-world cyber-physical applications and found that focusing only on the mean of reward-to-go and corresponding Bellman equation is impractical, especially in some safety-critical scenarios requiring tight risk control, such as autonomous vehicle navigation (Candela et al., 2023) and robot control (Zhao et al., 2021; Zhang et al., 2024). A primary reason is that the risk-neutral criterion (maximizing the expectation) ignores the characteristics of a reward distribution other than the mean, which may be important for systems with safety concerns. For example, a system may be required to operate in a manner that alleviates harmful consequences, even in rare situations that are difficult to predict.

To handle this kind of issue, some works (Wang & Chapman, 2022) introduce the worst-case criterion for autonomous systems with safety concerns to achieve zero-constraint violations by finding a policy that satisfies the constraints of a specific cost function, which generally assumes the maximum cost can quantify how bounded adversarial disturbances can inhibit the satisfactory operation of a system. However, due to the reliance on the typical assumption of bounded adversarial disturbances, the worst-case criterion may not be suitable for some applications that possess certain characteristics, such as the difficulty in characterizing the bounds of disturbances with a sufficient degree of certainty. Recently, risk-averse criterion (Bäuerle & Rieder, 2014; Bisi et al., 2022), an intermediary criterion between the risk-neutral and worst-case criteria, has garnered extensive attention, which describes people or algorithms that prefer outcomes with reduced uncertainty by seeking to optimize risk metrics, such as entropy risk measures (ERM) (Pichler & Schlotter, 2020) or conditional value-at-risk (CVaR) (Artzner, 1997; Chen et al., 2024), of the possible cumulative reward which emphasizes its distributional characteristics.

In general, there are mainly two types of risk-sensitive measures: nested and static quantile risk-aware measures, each possessing distinct advantages and limitations. Static risk measures (Fei et al., 2021; Wang et al., 2023) are straightforward to interpret, but the resulting optimal policy may not remain Markovian and may become history-dependent. On the other hand, nested risk measures (Chen et al., 2024; Zhao et al., 2024) utilize MDPs to ensure risk sensitivity of the value iteration at each step under the current state, resulting in a more conservative approach. In this paper, we prefer nested risk measures because they recursively adhere to the Bellman equation and allow the MDPs to be reconstructed through state augmentation, enabling them to remain Markovian and ensuring that policy choices depend solely on the current state.

Specifically, we introduce the popular CVaR (Artzner, 1997) objective as follows:

$$G(\xi) = \begin{cases} \frac{1}{\mu}\xi & \text{if } \xi < \mu, \\ 1 & \text{if } \xi \geq \mu, \end{cases} \tag{20}$$

and $\Phi^{\mu}(Z)$ becomes

$$\Phi^{\mu}(Z) = \frac{1}{\mu}\int_0^{\mu} F_Z^{-1}(\xi)\mathrm{d}\xi, \tag{21}$$

where G is L_G-Lipschitz continuous for some $L_G \in \mathbb{R}_{>0}$, and $G(0) = 0, G(1) = 1$.

Risks in LLMs Alignment. When aligning large language models with human preferences, there are many factors that may pose risks, primarily encompassing the following three types:

(1) There exist conflicts and contradictions among human preferences (or choices), thus introducing uncertainty in the objectives when aligning models with human preferences. In addition, human choice behavior has contextual choice effects (Peuter et al., 2024), i.e., a decision maker's choice between two options is influenced by adding more options to the choice set (Huber et al., 1982).

(2) Humans do not make decisions by maximizing their expected value for uncertain events; instead, they perceive random variables in a biased but well-defined manner (Ethayarajh et al., 2024). For example, relative to some reference point, humans are more sensitive to losses than gains, a phenomenon known as loss aversion.

(3) Many popular methods, such as DPO (Rafailov et al., 2023), RDPO (Fisch et al., 2024), and simPO (Meng et al., 2024), utilize KL divergence to ensure that the training model remains closely aligned with a reference model during the training process, preventing significant deviations. These methods still face the issue of being insensitive to strategic risks because

they only consider the mean of reward or utility and the corresponding Bellman equation, which is risk-neutral and does not capture the distribution characteristics of rewards efficiently.

Since the first two types of risks stem from the distribution of preference data itself, in this article, we focus on the third type of risk, which comes from the process during model alignment. Specifically, we investigate a novel direct preference optimization method for the problem of aligning with human preferences from a risk-sensitive perspective and provide theoretical and empirical results on its performance and risk-awareness.

A.2. The Expanded Version of Value Function Definition

The definition of value function for nested risk measure, i.e., Eq. 5 in Subsection 2.3, can be expanded as

$$\begin{aligned} Q_\pi\left(\left[x, y^{<t}\right], y^t\right)=R\left(\left[x, y^{<t}\right], y^t\right)+\Phi^\mu\left(R\left(\left[x, y^{<t+1}\right], \pi\left(\cdot \mid\left[x, y^{<t+1}\right]\right)\right)\right. \\ \left.+\Phi^\mu\left(\cdots \Phi^\mu\left(R\left(\left[x, y^{<T}\right], \pi\left(\cdot \mid\left[x, y^{<T}\right]\right)\right)\right)\right)\right), \end{aligned} \tag{22}$$

$$\begin{aligned} V_\pi\left(\left[x, y^{<t}\right]\right)=R\left(\left[x, y^{<t}\right], \pi\left(\cdot \mid\left[x, y^{<t}\right]\right)\right)+\Phi^\mu\left(R\left(\left[x, y^{<t+1}\right], \pi\left(\cdot \mid\left[x, y^{<t+1}\right]\right)\right)\right. \\ \left.+\Phi^\mu\left(\cdots \Phi^\mu\left(R\left(\left[x, y^{<T}\right], \pi\left(\cdot \mid\left[x, y^{<T}\right]\right)\right)\right)\right)\right). \end{aligned} \tag{23}$$

Similarly, the definition of the optimal value function, can be expanded as

$$\begin{aligned} Q^*_\pi\left(\left[x, y^{<t}\right], y^t\right)=\max\left\{R\left(\left[x, y^{<t}\right], y^t\right)+\Phi^\mu\left(R\left(\left[x, y^{<t+1}\right], \pi\left(\cdot \mid\left[x, y^{<t+1}\right]\right)\right)\right.\right. \\ \left.\left.+\Phi^\mu\left(\cdots \Phi^\mu\left(R\left(\left[x, y^{<T}\right], \pi\left(\cdot \mid\left[x, y^{<T}\right]\right)\right)\right)\right)\right)\right\}, \end{aligned} \tag{24}$$

$$\begin{aligned} V^*_\pi\left(\left[x, y^{<t}\right]\right)=\max\left\{R\left(\left[x, y^{<t}\right], \pi\left(\cdot \mid\left[x, y^{<t}\right]\right)\right)+\Phi^\mu\left(R\left(\left[x, y^{<t+1}\right], \pi\left(\cdot \mid\left[x, y^{<t+1}\right]\right)\right)\right.\right. \\ \left.\left.+\Phi^\mu\left(\cdots \Phi^\mu\left(R\left(\left[x, y^{<T}\right], \pi\left(\cdot \mid\left[x, y^{<T}\right]\right)\right)\right)\right)\right)\right\}. \end{aligned} \tag{25}$$

B. Supplementary Materials for Section 3

B.1. The Proof of Lamma 3.1

Lemma 3.2 Restated. For a given Pb-MDP, the reward on the entire prompt-response can be decomposed as $r = \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right)$, $V_\pi[x]$ in Eq. 5 and $\tilde{V}_\pi[x]$ in Eq. 6 are equivalent, which implies the following characteristics:

Proof. Firstly, according to (Givan et al., 2003; Lowd & Davis, 2010; Zhao et al., 2024), we can reformulate the Pb-MDP as a decision tree-like MDP.

(1) The state transition graph of the Pb-MDP is connected and acyclic;

(2) Each state in the Pb-MDP corresponds to a unique node in the tree;

(3) There is a single root node from which every other node is reachable via a unique path;

(4) The transition probabilities between states follow the Markov property, i.e., the probability of transitioning to any future state depends only on the current state and not on the sequence of events that preceded it.

Formally, let S be the set of states and p_{ij} be the transition probabilities between states $\mathbf{s}_i$ and $\mathbf{s}_j$. For an Pb-MDP with a tree-like structure, the probabilistic transition matrix P is defined such that:

$$p_{ij} > 0 \text{ if there is an edge between } \mathbf{s}_i \text{ and } \mathbf{s}_j \text{ in the tree, and } p_{ij} = 0 \text{ otherwise.} \tag{26}$$

Moreover, for each non-root node $\mathbf{s}_j$, there exists exactly one $\mathbf{s}_i$ such that $p_{ij} > 0$, and $\mathbf{s}_i$ is the unique parent of $\mathbf{s}_j$ in the tree structure.

To classify the two value iteration in Eq. 5 and Eq. 6, we denote the value given by Eq. 6 as $\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)$ and the value given by Eq. 5 as $V_\pi\left(\left[x, y^{<t}\right]\right)$, thus, in tree-like Pb-MDP with the reward of the entire prompt-response, which can be decomposed as $r = \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right)$, we have the following relationship:

$$\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right) = V_\pi\left(\left[x, y^{<t}\right]\right) + R_{1:t-1},$$

where $R_{1:t-1} = \sum_{h=1}^{t-1} \gamma^{h-1} R\left(\left[x, y^{<h}\right], y^h\right)$ denotes the reward of the $1 \sim t-1$ steps of a prompt-response. We prove this relationship by mathematical induction.

Initial Case. Using the tree-like Pb-MDP and the initial conditions of the Bellman equation, at the final step $t = T$, we have

$$\begin{aligned} \tilde{V}_\pi\left(\left[x, y^{<T}\right]\right) &= V_\pi\left(\left[x, y^{<T}\right], \pi\left(\cdot \mid \left[x, y^{<t}\right]\right)\right) + R_{1:T-1} \\ &= V_\pi\left(\left[x, y^{<T}\right]\right) + R_{1:T-1}. \end{aligned} \tag{27}$$

Induction Step. We now proved that if $\tilde{V}_\pi\left(\left[x, y^{<t+1}\right]\right) = V_\pi\left(\left[x, y^{<t+1}\right]\right) + R_{1:t}$ holds, then $\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right) = V_\pi\left(\left[x, y^{<t}\right]\right) + R_{1:t-1}$ also holds. Since this policy π on tree-like Pb-MDP is fixed, it has only one path to arrive t-th state ($s_t = [x, y^{<t}]$), denoted as:

$$\Xi_t\left(s_{T,1}\right) = \Xi_h\left(s_{T,2}\right) \quad \forall\, s_{T,1}, s_{T,2} \in \left\{s_T \mid S_t\left(s_T\right) = \left[x, y^{<t}\right]\right\}.$$

Therefore, $R_{1:t-1}$ is unique.

$$\begin{aligned} \tilde{V}_\pi\left(\left[x, y^{<t}\right]\right) &= \Phi^\mu\left(V_\pi\left(\left[x, y^{<t+1}\right]\right) + R_{1:t}\right), \\ &= \Phi^\mu\left(V_\pi\left(\left[x, y^{<t+1}\right]\right) + R\left(\left[x, y^{<t}\right], \pi\left(\cdot \mid \left[x, y^{<t}\right]\right)\right) + R_{1:t-1}\right), \\ &= \Phi^\mu\left(V_\pi\left(\left[x, y^{<t+1}\right]\right) + R\left(\left[x, y^{<t}\right], \pi\left(\cdot \mid \left[x, y^{<t}\right]\right)\right)\right) + R_{1:t-1}, \\ &= V_\pi\left(\left[x, y^{<t+1}\right]\right) + R_{1:t-1}, \end{aligned} \tag{28}$$

where the third equality holds because the risk measure function Φ satisfies translation invariance. Then, by applying conclusion, we observe that when $t = 1$, $\tilde{V}_\pi[x] = V_\pi[x]$ hold on. Thus, we have proven that for the Pb-MDP, the reward of the entire trajectory can be decomposed as $r = \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right)$, and $V_\pi[x]$ in Eq. 5 and $\tilde{V}_\pi[x]$ in Eq. 6 are equivalent. □

B.2. The derivation of Definition 3.2

Definition 3.3 Restated. For a risk-sensitive Pb-MDP that satisfies the Bellman equation in Eq. 6, the risk-aware advantage function can be defined as

$$\tilde{A}_\pi\left(\left[x, y^{<t}\right], z\right) = \tilde{Q}_\pi\left(\left[x, y^{<t}\right], z\right) - \Phi^\mu(\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)),$$

where z subject to $\pi_\theta\left(\cdot \mid \left[x, y^{<t}\right]\right)$.

In terms of designing the objective function at the token level, (Zeng et al., 2024) provides us with a valuable insight by introducing the advantage function from the TRPO algorithm in reinforcement learning as the target for each step. In this paper, building upon TDPO, we consider the risk associated with language generation at each step and devise a novel risk-sensitive advantage function. First, based on assumption that $r = \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right)$, we can get:

$$\begin{aligned} r &= \sum_{t=1}^{T} \gamma^{t-1} R\left(\left[x, y^{<t}\right], y^t\right) \\ &= \sum_{t=1}^{T} \gamma^{t-1} \left(R\left(\left[x, y^{<t}\right], y^t\right) + \gamma\, \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t+1}\right]\right)\right) - \gamma\, \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t+1}\right]\right)\right)\right) \\ &= \Phi^\mu\left(\tilde{V}_\pi\left([x]\right)\right) + \sum_{t=1}^{T} \gamma^{t-1} \left(R\left(\left[x, y^{<t}\right], y^t\right) + \gamma\, \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t+1}\right]\right)\right) - \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)\right)\right) - \gamma^T\, \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<T+1}\right]\right)\right) \\ &= \Phi^\mu\left(\tilde{V}_\pi\left([x]\right)\right) + \sum_{t=1}^{T} \gamma^{t-1} \left(\tilde{Q}_\pi\left(\left[x, y^{<t}\right], y^t\right) - \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)\right)\right) - \gamma^T\, \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<T+1}\right]\right)\right). \end{aligned} \tag{29}$$

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

$$V_\pi\left(\left[x, y^{<T+1}\right]\right) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R\left(\left[x, y^{<T+1+k}\right], y^{T+1+k}\right) \mid s_t = \left[x, y^{<T+1}\right]\right] = 0. \tag{30}$$

Furthermore, we have

$$r = \Phi^{\mu}\left(\tilde{V}_{\pi}\left([x]\right)\right) + \sum_{t=1}^{T} \gamma^{t-1}\left(\tilde{Q}_{\pi}\left(\left[x, y^{<t}\right], y^{t}\right) - \Phi^{\mu}\left(\tilde{V}_{\pi}\left(\left[x, y^{<t}\right]\right)\right)\right). \tag{31}$$

So, we definite the risk-aware advantage function as $\tilde{A}_{\pi}\left(\left[x, y^{<t}\right], z\right) = \tilde{Q}_{\pi}\left(\left[x, y^{<t}\right], z\right) - \Phi^{\mu}\left(\tilde{V}_{\pi}\left(\left[x, y^{<t}\right]\right)\right)$, where $z \sim \pi_{\theta}\left(\cdot \mid [x, y^{<t}]\right)$.

B.3. The Proof of Lemma 3.3

Lemma 3.4 Restated. Given two policies π and π', if for any state $s_t = \left[x, y^{<t}\right], \mathbb{E}_{z\sim\pi'}\left[\tilde{A}_{\pi}\left(\left[x, y^{<t}\right], z\right)\right] \geq 0$ holds, then we can conclude:

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

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

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{V}_{\pi'}([x])\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{V}_{\pi}([x])\right] \\
=&\mathbb{E}_{\tau|\pi'}\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(R\left(\left[x, y^{<t}\right], y^{t}\right) + \gamma\,\Phi^{\mu}\left(\tilde{V}_{\pi}\left(\left[x, y^{<t+1}\right]\right)\right)\right) - \tilde{V}_{\pi}([x])\right] \\
=&\mathbb{E}_{\tau|\pi'}\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(R\left(\left[x, y^{<t}\right], y^{t}\right) + \gamma\,\Phi^{\mu}\left(\tilde{V}_{\pi}\left(\left[x, y^{<t+1}\right]\right)\right) - \Phi^{\mu}\left(\tilde{V}_{\pi}\left(\left[x, y^{<t}\right]\right)\right)\right)\right] \\
=&\mathbb{E}_{\tau|\pi'}\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(\tilde{A}_{\pi}\left(\left[x, y^{<t}\right], y^{t}\right)\right)\right] \\
=&\mathbb{E}_{\tau|\pi'}\left[\sum_{t=1}^{\infty}\gamma^{t-1}\left(\mathbb{E}_{y^t\sim\pi'}\left[\tilde{A}_{\pi}\left(\left[x, y^{<t}\right], y^{t}\right)\right)\right]\right].
\end{aligned} \tag{32}$$

Since for any state $s_t = \left[x, y^{<t}\right], \mathbb{E}_{z\sim\pi'}\left[\tilde{A}_{\pi}\left(\left[x, y^{<t}\right], z\right)\right] \geq 0$, so we can obtain

$$\mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{V}_{\pi'}([x])\right] - \mathbb{E}_{x\sim\mathcal{D}}\left[\tilde{V}_{\pi}([x])\right] \geq 0.$$

□

B.4. The Proof of Lemma 3.4

Lemma 3.5 Restated. The constrained problem in Eq. 9 has the closed-form solution:

$$\pi_{\theta}^{*}\left(z \mid \left[x, y^{<t}\right]\right) = \frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right)\exp\left(\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right)\right)}{Z\left(\left[x, y^{<t}\right]; \beta\right)},$$

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

Proof.

$$\begin{aligned}
&\max_{\pi_\theta} \mathbb{E}_{z\sim\pi_\theta(\cdot|[x,y^{<t}])} \tilde{A}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right) - \beta D_{\text{KL}}\left(\pi_\theta\left(\cdot \mid \left[x, y^{<t}\right]\right) \| \pi_{\text{ref}}\left(\cdot \mid \left[x, y^{<t}\right]\right)\right) \\
=&\max_{\pi_\theta} \mathbb{E}_{z\sim\pi_\theta(\cdot|[x,y^{<t}])} \left(\left(\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right) - \tilde{V}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right]\right)\right) + \beta \log\left(\frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right)}{\pi_\theta\left(z \mid [x, y^{<t}]\right)}\right)\right) \\
=&\max_{\pi_\theta} \beta \mathbb{E}_{z\sim\pi_\theta(\cdot|[x,y^{<t}])} \log\left(\frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right) e^{\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x,y^{<t}\right],z\right)}}{\pi_\theta\left(z \mid [x, y^{<t}]\right)}\right) - \tilde{V}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right]\right) \\
=&\max_{\pi_\theta} \beta \mathbb{E}_{z\sim\pi_\theta(\cdot|[x,y^{<t}])} \log\left(\frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right) e^{\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x,y^{<t}\right],z\right)}}{Z\left([x, y^{<t}]; \beta\right)\pi_\theta\left(z \mid [x, y^{<t}]\right)}\right) \\
&- \tilde{V}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right]\right) + \beta \log Z\left(\left[x, y^{<t}\right]; \beta\right) \\
=&\max_{\pi_\theta} -\beta D_{\text{KL}}\left(\pi_\theta\left(z \mid \left[x, y^{<t}\right]\right) \| \frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right) e^{\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x,y^{<t}\right],z\right)}}{Z\left([x, y^{<t}]; \beta\right)}\right) \\
&- \tilde{V}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right]\right) + \beta \log Z\left(\left[x, y^{<t}\right]; \beta\right),
\end{aligned} \tag{33}$$

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

$$Z\left(\left[x, y^{<t}\right]; \beta\right) = \mathbb{E}_{z\sim\pi_{\text{ref}}(\cdot|[x,y^{<t}])} \exp\left(\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], z\right)\right). \tag{34}$$

Then, we can derive the relationship between the optimal policy and the state-action function:

$$\pi_\theta^*\left(z \mid \left[x, y^{<t}\right]\right) = \frac{\pi_{\text{ref}}\left(z \mid [x, y^{<t}]\right) \exp\left(\frac{1}{\beta}\tilde{Q}_{\pi_{\text{ref}}}\left([x, y^{<t}], z\right)\right)}{Z\left([x, y^{<t}]; \beta\right)}. \tag{35}$$

□

B.5. The Proof of Lemma 3.5

Lemma 3.6 Restated. Given a reward function r, based on a relationship between token-wise rewards and the reward function represented by $r = \sum_{t=1}^{T} \gamma^{t-1} R\left([x, y^{<t}], y^t\right)$, we can establish the equivalence between the Bradley-Terry model and the Regret Preference Model in the language generation task, i.e.,

$$P_{\text{BT}}\left(y_1 \succ y_2 \mid x\right) = \sigma\left(\sum_{t=1}^{T_1} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_1^{<t}\right], y_1^t\right) - \sum_{t=1}^{T_2} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_2^{<t}\right], y_2^t\right)\right), \tag{36}$$

where $\sigma(z) = 1/\left(1 + \exp(-z)\right)$ is the logistic sigmoid function for any random variable z.

Proof. Recalling to the BT model in Eq. 40

$$P_{\text{BT}}\left(y_1 \succ y_2 \mid x\right) = \frac{\exp\left(r\left(x, y_1\right)\right)}{\exp\left(r\left(x, y_1\right)\right) + \exp\left(r\left(x, y_2\right)\right)}, \tag{37}$$

and the equivalence between prompt-response reward and the risk-aware advantage function:

$$\begin{aligned}
r &= \Phi^\mu\left(\tilde{V}_\pi\left([x]\right)\right) + \sum_{t=1}^{T} \gamma^{t-1}\left(\tilde{Q}_\pi\left(\left[x, y^{<t}\right], y^t\right) - \Phi^\mu\left(\tilde{V}_\pi\left(\left[x, y^{<t}\right]\right)\right)\right) \\
&= \Phi^\mu\left(\tilde{V}_\pi\left([x]\right)\right) + \sum_{t=1}^{T} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y^{<t}\right], y^t\right).
\end{aligned}$$

Then, we have

$$P_{\text{BT}}\left(y_1 \succ y_2 \mid x\right) = \sigma\left(\sum_{t=1}^{T_1} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_1^{<t}\right], y_1^t\right) - \sum_{t=1}^{T_2} \gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_2^{<t}\right], y_2^t\right)\right).$$

□

B.6. The Proof of Theorem 3.6

Theorem 3.7 Restated. Given prompts x and pairwise responses (y_1, y_2), and the risk-aware objective function in Eq. 9, the Bradley-Terry model expresses the human preference probability in terms of the risk-aware optimal policy π_θ^* and reference policy π_{ref}:

$$P_{\mathrm{BT}}^*\left(y_1 \succ y_2 \mid x\right)=\sigma\left(u^*\left(x, y_1, y_2\right)-\delta^*\left(x, y_1, y_2\right)\right),$$

where $u\left(x, y_1, y_2\right)$ represents the difference in implicit rewards defined by the risk-aware policy π_θ^* and the reference policy π_{ref}, weighted by β, represented as

$$u\left(x, y_1, y_2\right)=\beta \log \frac{\pi_\theta\left(y_1 \mid x\right)}{\pi_{\mathrm{ref}}\left(y_1 \mid x\right)}-\beta \log \frac{\pi_\theta\left(y_2 \mid x\right)}{\pi_{\mathrm{ref}}\left(y_2 \mid x\right)},$$

and $\delta\left(x, y_1, y_2\right)$ represents the difference in sequential risk ratio between two pairs (x, y_1) and (x, y_2), expressed as

$$\delta\left(x, y_1, y_2\right)=\beta D_{\mathrm{SeqRR}}\left(x, y_2 ; \pi_{\mathrm{ref}} \mid \pi_\theta\right)-\beta D_{\mathrm{SeqRR}}\left(x, y_1 ; \pi_{\mathrm{ref}} \mid \pi_\theta\right).$$

Proof. According to the Lemma 3.4, we have

$$\pi_\theta^*\left(z \mid \left[x, y^{<t}\right]\right)=\frac{\pi_{\mathrm{ref}}\left(z \mid\left[x, y^{<t}\right]\right) \exp \left(\frac{1}{\beta} \tilde{Q}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], z\right)\right)}{Z\left(\left[x, y^{<t}\right] ; \beta\right)}, \tag{38}$$

where $Z\left(\left[x, y^{<t}\right] ; \beta\right)=\mathbb{E}_{z \sim \pi_{\mathrm{ref}}\left(\cdot \mid\left[x, y^{<t}\right]\right)} e^{\frac{1}{\beta} \tilde{Q}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], z\right)}$ is the partition function. Rearrange Eq. 38, we obtain

$$\tilde{Q}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], z\right)=\beta \log \frac{\pi_\theta^*\left(z \mid\left[x, y^{<t}\right]\right)}{\pi_{\mathrm{ref}}\left(z \mid\left[x, y^{<t}\right]\right)}+\beta \log Z\left(\left[x, y^{<t}\right] ; \beta\right). \tag{39}$$

From Lemma 3.5, we can get

$$P_{\mathrm{BT}}\left(y_1 \succ y_2 \mid x\right)=\sigma\left(\sum_{t=1}^{T_1}\left(\gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_1^{<t}\right], y_1^t\right)\right)-\sum_{t=1}^{T_2}\left(\gamma^{t-1} \tilde{A}_\pi\left(\left[x, y_2^{<t}\right], y_2^t\right)\right)\right). \tag{40}$$

By leveraging Eq. 39, we can derive

$$\begin{aligned}
&\sum_{t=1}^T \gamma^{t-1} \tilde{A}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], y^t\right) \\
=&\sum_{t=1}^T \gamma^{t-1}\left(Q_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], y^t\right)-\Phi^\mu\left(\tilde{V}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right]\right)\right)\right) \\
=&\sum_{t=1}^T \gamma^{t-1}\left(\tilde{Q}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], y^t\right)-\Phi^\mu\left(\tilde{Q}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], z\right)\right)\right) \\
=&\sum_{t=1}^T \gamma^{t-1}\left(\beta \log \frac{\pi_\theta^*\left(y^t \mid\left[x, y^{<t}\right]\right)}{\pi^{\mathrm{ref}}\left(y^t \mid\left[x, y^{<t}\right]\right)}+\beta \log Z\left(\left[x, y^{<t}\right] ; \beta\right)-\Phi^\mu\left(\beta \log \frac{\pi_\theta^*\left(z \mid\left[x, y^{<t}\right]\right)}{\pi^{\mathrm{ref}}\left(z \mid\left[x, y^{<t}\right]\right)}+\beta \log Z\left(\left[x, y^{<t}\right] ; \beta\right)\right)\right).
\end{aligned} \tag{41}$$

Note that

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

Therefore,

$$\begin{aligned}
&\sum_{t=1}^T \gamma^{t-1} \tilde{A}_{\pi_{\mathrm{ref}}}\left(\left[x, y^{<t}\right], y^t\right) \\
=&\beta \sum_{t=1}^T \gamma^{t-1}\left(\log \frac{\pi_\theta^*\left(y^t \mid\left[x, y^{<t}\right]\right)}{\pi_{\mathrm{ref}}\left(y^t \mid\left[x, y^{<t}\right]\right)}-\Phi_{z \sim \pi_{\mathrm{ref}}}^\mu\left(\log \frac{\pi_\theta^*\left(z \mid\left[x, y^{<t}\right]\right)}{\pi_{\mathrm{ref}}\left(z \mid\left[x, y^{<t}\right]\right)}\right)\right) \\
=&\beta \sum_{t=1}^T \gamma^{t-1} \log \frac{\pi_\theta^*\left(y^t \mid\left[x, y^{<t}\right]\right)}{\pi_{\mathrm{ref}}\left(y^t \mid\left[x, y^{<t}\right]\right)}+\beta \sum_{t=1}^T \gamma^{t-1} \Phi_{z \sim \pi_{\mathrm{ref}}}^\mu\left(\log \frac{\pi_\theta^*\left(z \mid\left[x, y^{<t}\right]\right)}{\pi_{\mathrm{ref}}\left(z \mid\left[x, y^{<t}\right]\right)}\right).
\end{aligned} \tag{42}$$

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

$$\begin{aligned}\sum_{t=1}^{T} \tilde{A}_{\pi_{\text{ref}}}\left(\left[x, y^{<t}\right], y^{t}\right) =& \beta \sum_{t=1}^{T} \log \frac{\pi_{\theta}^{*}\left(y^{t} \mid\left[x, y^{<t}\right]\right)}{\pi_{\text{ref}}\left(y^{t} \mid\left[x, y^{<t}\right]\right)}+\beta \sum_{t=1}^{T} \Phi_{z \sim \pi_{\text{ref}}}^{\mu}\left(\log \frac{\pi_{\theta}^{*}\left(z \mid\left[x, y^{<t}\right]\right)}{\pi_{\text{ref}}\left(z \mid\left[x, y^{<t}\right]\right)}\right) \\ =& \beta\left(\log \frac{\pi_{\theta}^{*}(y \mid x)}{\pi_{\text{ref}}(y \mid x)}+D_{\text{SeqRR}}\left(x, y ; \pi_{\text{ref}} \mid \pi_{\theta}^{*}\right)\right),\end{aligned} \tag{43}$$

where $D_{\text{SeqRR}}\left(x, y ; \pi_{\text{ref}} \mid \pi_{\theta}\right)=\sum_{t=1}^{T} \Phi_{z \sim \pi_{\text{ref}}}^{\mu}\left(\log \frac{\pi_{\text{ref}}(z \mid x)}{\pi_{\theta}(z \mid x)}\right)$.

Then, we let

$$u\left(x, y_{1}, y_{2}\right)=\beta \log \frac{\pi_{\theta}\left(y_{1} \mid x\right)}{\pi_{\text{ref}}\left(y_{1} \mid x\right)}-\beta \log \frac{\pi_{\theta}\left(y_{2} \mid x\right)}{\pi_{\text{ref}}\left(y_{2} \mid x\right)}, \tag{44}$$

$$\delta\left(x, y_{1}, y_{2}\right)=\beta D_{\text{SeqRR}}\left(x, y_{2} ; \pi_{\text{ref}} \mid \pi_{\theta}\right)-\beta D_{\text{SeqRR}}\left(x, y_{1} ; \pi_{\text{ref}} \mid \pi_{\theta}\right). \tag{45}$$

Substituting Eq. 43 into Eq. 40, we arrive at $P_{\text{BT}}^{*}\left(y_{1} \succ y_{2} \mid x\right)=\sigma\left(u^{*}\left(x, y_{1}, y_{2}\right)-\delta^{*}\left(x, y_{1}, y_{2}\right)\right)$.

□

B.7. Algorithm

In this subsection, we provide the main pseudocode for Risk-aware Direct Preference Optimization (Ra-DPO), as outlined in Algorithm 1.

Algorithm 1 Risk-aware Direct Preference Optimization (Ra-DPO)

Input: Reference model π_{ref}, Policy model π_θ, Coefficient α, β, Risk control parameter μ, Learning rate η
Input: Dataset $\mathcal{D}=\left\{\left(x, y_{w}, y_{l}\right)^{i}\right\}_{i=1}^{N}$ of size N, Method $\mathcal{M}$
Initialize: $\pi_\theta \leftarrow \pi_{\text{ref}}$
for each epoch **do**
  Sample mini-batch $\mathcal{D}_m=\left\{\left(x, y_{w}, y_{l}\right)^{m}\right\}_{m=1}^{M}$ from $\mathcal{D}$
  Predict the probabilities $\pi_\theta(y_w \mid x)$ and $\pi_\theta(y_l \mid x)$ for (x, y_w, y_l) in the mini-batch $\mathcal{D}_m$ using the policy model
  Predict the probabilities $\pi_{\text{ref}}(y_w \mid x)$ and $\pi_{\text{ref}}(y_l \mid x)$ for (x, y_w, y_l) in the mini-batch $\mathcal{D}_m$ using the reference model
  Calculate the function $u(x, y_w, y_l)=\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}$
  Compute the sequential risk ratio $D_{\text{SeqRR}}(x, y_w; \pi_{\text{ref}} \mid \pi_\theta)$ for (x, y_w) in the mini-batch $\mathcal{D}_m$
  Compute the sequential risk ratio $D_{\text{SeqRR}}(x, y_l; \pi_{\text{ref}} \mid \pi_\theta)$ for (x, y_l) in the mini-batch $\mathcal{D}_m$
  if Method $\mathcal{M}$ is Ra-DPO$_1$ **then**
    Calculate the function $\delta(x, y_w, y_l)=\beta D_{\text{SeqRR}}(x, y_l; \pi_{\text{ref}} \mid \pi_\theta)-\beta D_{\text{SeqRR}}(x, y_w; \pi_{\text{ref}} \mid \pi_\theta)$
    $\theta \leftarrow \theta+\eta \nabla_\theta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_m}\left[\log \sigma\left(u(x, y_w, y_l)-\delta(x, y_w, y_l)\right)\right]$
  else {Method $\mathcal{M}$ is Ra-DPO$_2$}
    Calculate the function $\delta_2(x, y_w, y_l)=\beta D_{\text{SeqRR}}(x, y_l; \pi_{\text{ref}} \mid \pi_\theta)-\text{sg}\left(\beta D_{\text{SeqRR}}(x, y_w; \pi_{\text{ref}} \mid \pi_\theta)\right)$
    $\theta \leftarrow \theta+\eta \nabla_\theta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_m}\left[\log \sigma\left(u(x, y_w, y_l)-\alpha\delta_2(x, y_w, y_l)\right)\right]$
  end if
end for

C. Supplementary Materials for Section 4

C.1. Additional experimental results

In this paper, we evaluate the performance of our proposed algorithm, Ra-DPO (Algorithm 1 in the Appendix B.7), against baseline algorithms on several text tasks. Here, we provide some additional experimental results, which are illustrated in Figures 6-7.

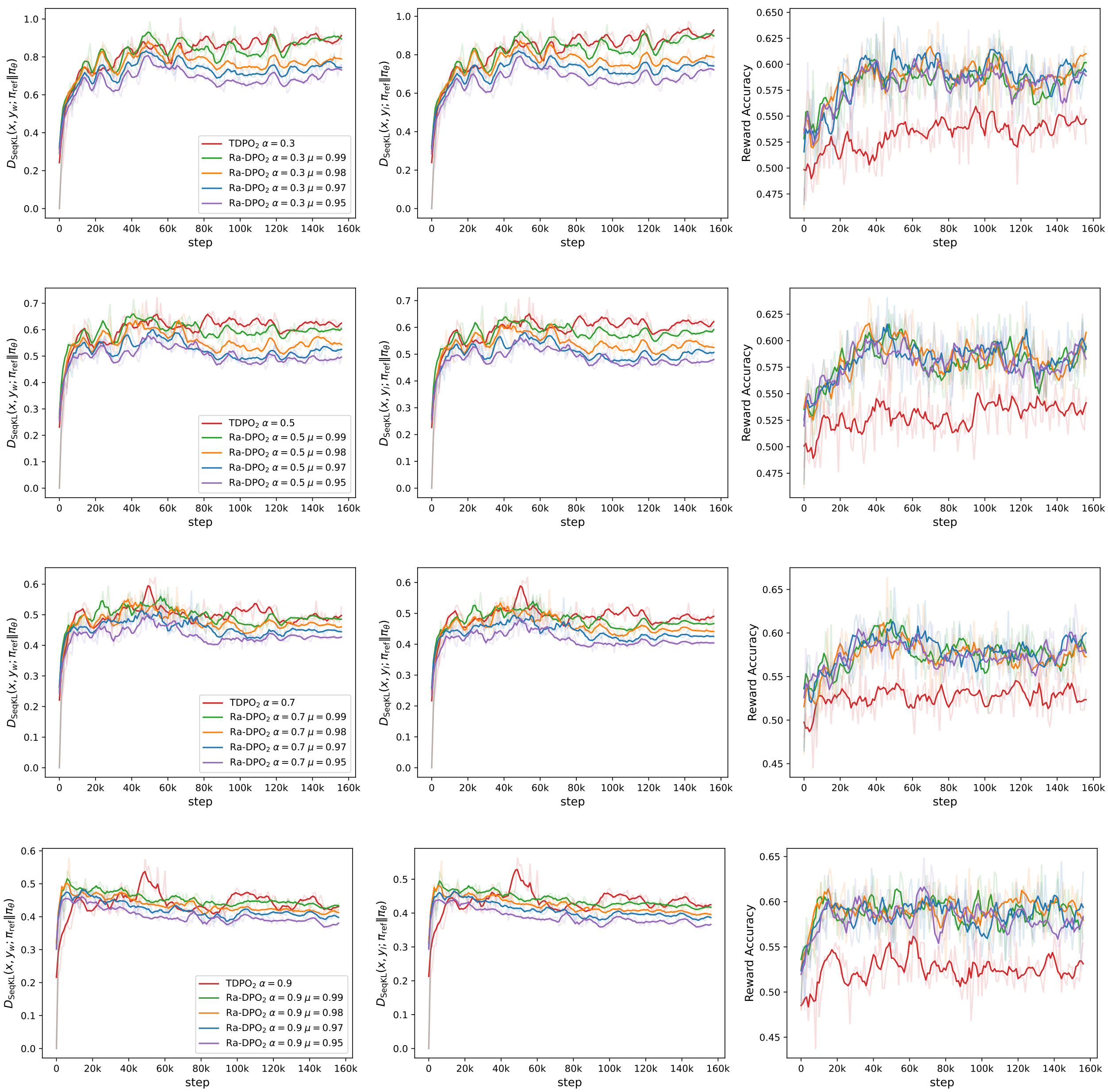


Figure 6. The experiment on the Anthropic HH dataset with Pythia-1.4B serving as the base model. We implemented TDPO$_2$, and different versions of Ra-DPO$_2$ with respect to the parameters α and μ. The progression of sequential KL divergence (the lower the better) of both preferred response and dispreferred responses are presented on the left and in the middle. Additionally, the reward accuracy curves (the higher the better) are illustrated on the right.

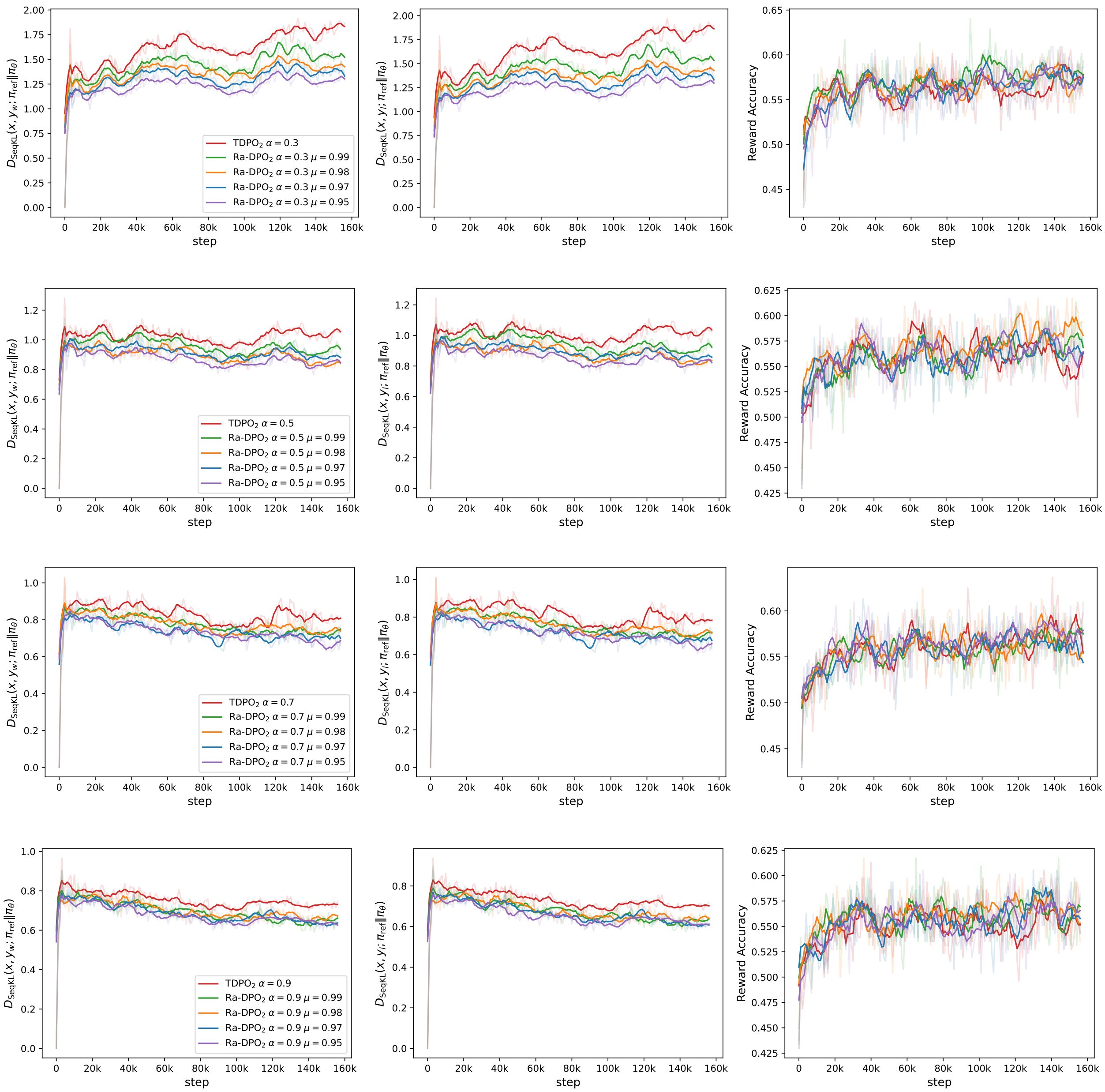


Figure 7. The experiment on the Anthropic HH dataset with Pythia-2.8B serving as the base model. We implemented TDPO$_2$, and different versions of Ra-DPO$_2$ with respect to the parameters α and μ. The progression of sequential KL divergence (the lower the better) of both preferred response and dispreferred responses are presented on the left and in the middle. Additionally, the reward accuracy curves (the higher the better) are illustrated on the right.